Echantillonage préférentiel pour l'approximation de gradients dans une famille exponentielle naturelle discrète

Bastien Batardière, Julien Chiquet, Joon Kwon and Julien Stoehr

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, Université Paris-Dauphine

March 24, 2024

Outline

Introduction

Model and assumptions

Algorithm

Theorem with biased estimator

Bound on the bias

Method and some guarantees

Latent variable models

- Latent variable model: $\boldsymbol{Y}_i \in \mathbb{R}^p$ is driven by a latent variable $\boldsymbol{W} \in \mathbb{R}^q$:

$$p_\theta(\boldsymbol{Y}_i) = \int_{\mathbb{R}^q} p_\theta(Y_i, \boldsymbol{W}) \mathrm{d}\boldsymbol{W}$$

  with a parameter $\theta \in \mathbb{R}^d$ and $1 \leq i \leq n$ with n the number of samples.
- PLN with $\boldsymbol{Y}_i | \boldsymbol{W} \sim \mathcal{P}(\exp(\boldsymbol{W})), \quad \boldsymbol{W} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- PLN-PCA with $\boldsymbol{Y}_i | \boldsymbol{W} \sim \mathcal{P}(\exp(C\boldsymbol{W} + \boldsymbol{\mu}), \quad W \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{I}_q)$
- Multivariate Binomial, Mixture models ...

Natural exponential family

We assume that p_θ belongs to the natural exponential family and the dependance in θ is linear:

$$\mathbf{W}_i \sim^{\text{iid}} \mathcal{N}(\mathbf{0}_q, \mathbf{I}_q), \quad \mathbf{Z}_i = \mathbf{C}\mathbf{W}_i + \boldsymbol{\mu},$$

$$p_\theta\left(Y_{ij}|Z_{ij}\right) = \exp(Y_{ij}Z_{ij} - A(Z_{ij}) - h(Y_{ij})), \quad 1 \leq j \leq p\,,$$

where h and A are real-valued functions with A convex and differentiable, $q << p$ and $\theta = (\boldsymbol{C}, \boldsymbol{\mu})$.

Goal and assumptions

- Goal: maximize the (non-concave) log-likelihood

$$\operatorname*{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(\boldsymbol{Y}_i) = \operatorname*{argmax}_{\theta} \ell(\theta).$$

- We first assume that $\theta \mapsto \log p_\theta(Y_i)$ is $\mathcal{C}^1$ (condition satisfied in the Poisson and Binomial case).

Algorithm

- Given a learning rate $\eta > 0$ and $\theta^{(0)} \in \mathbb{R}^d$ an initial point, we recursively define $\theta^{(t)}$ via Stochastic Gradient Ascent:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \widehat{g}^{(t)}$$

  where $\widehat{g}^{(t)}$ is a (possibly biased) gradient estimator of $\nabla \ell(\theta^{(t)})$.

Gradient estimator

We are given a family of law $\pi(\cdot; \theta, i)$ for $1 \leq i \leq n$ and $\theta \in \mathbb{R}^d$.

- At iteration $t \geq 1$, an index $i(t) \sim \text{Unif}\{1, \ldots, n\}$ is sampled and the law
$$\pi^{(t)} \triangleq \pi(\cdot; \theta^{(t)}, i(t))$$
is selected.
- Monte Carlo particles are sampled $(\mathbf{V}_k)_{1 \leq k \leq N} \overset{\text{iid}}{\sim} \pi^{(t)}$, with $N \geq 1$ a fixed number of particles.
- A self normalized gradient estimator is computed:
$$\widehat{g}^{(t)} \triangleq \textstyle\sum_{k=1}^{N} \omega_k \nabla_\theta \log p_{\theta^{(t)}}(\mathbf{Y}_{i(t)}, \mathbf{V}_k),$$
$$\omega_k = \frac{\rho_k}{\sum_{\ell=1}^{N} \rho_\ell} \text{ with } \rho_k = \frac{p_{\theta^{(t)}}(\mathbf{Y}_{i(t)}, \mathbf{V}_k)}{\pi^{(t)}(\mathbf{V}_k)}.$$

Gradient formula

Why such a gradient estimator ?

$$
\begin{aligned}
\nabla \log p_\theta(\boldsymbol{Y}_i) &= \mathbb{E}_{\boldsymbol{W}|\boldsymbol{Y}_i}\big[\overbrace{\nabla \log p_\theta(\boldsymbol{Y}_i|\boldsymbol{W})}^{h_i(\boldsymbol{W})}\big] \\
&= \mathbb{E}_{\boldsymbol{W}|\boldsymbol{Y}_i}\left[h_i(\boldsymbol{W})\right] \\
&= \mathbb{E}_\pi\left[\frac{p_\theta(\boldsymbol{V}|\boldsymbol{Y}_i)}{\pi(\boldsymbol{V})}h_i(\boldsymbol{V})\right] \\
&= \frac{1}{p_\theta(\boldsymbol{Y}_i)}\mathbb{E}_\pi\left[\frac{p_\theta(\boldsymbol{V},\boldsymbol{Y}_i)}{\pi(\boldsymbol{V})}h_i(\boldsymbol{V})\right] \\
&\overset{LLN}{\approx} \frac{1}{p_\theta(\boldsymbol{Y}_i)}\frac{1}{N}\sum_{k=1}^{N}\frac{p_\theta(\boldsymbol{V}_k,\boldsymbol{Y}_i)}{\pi(\boldsymbol{V}_k)}h_i(\boldsymbol{V}_k)
\end{aligned}
$$

with $\boldsymbol{V}_k \sim \pi$. The $p_\theta(\boldsymbol{Y}_i)$ term is unknown and estimated via IS:

$$
p_\theta(\boldsymbol{Y}_i) = \mathbb{E}_{\boldsymbol{W}}[p_\theta(\boldsymbol{Y}_i|\boldsymbol{W})] = \mathbb{E}_\pi\left[\frac{p_\theta(\boldsymbol{Y}_i|\boldsymbol{W})}{\pi(\boldsymbol{W})}p(\boldsymbol{W})\right] \approx \frac{1}{N}\sum_{k=1}^{N}\frac{p_\theta(\boldsymbol{V}_k,\boldsymbol{Y}_i)}{\pi(\boldsymbol{V}_k)}
$$

Pseudo-code

Algorithm 1: Pseudo code SGIS

Input $\theta^{(0)} \in \mathbb{R}^d$ initial point, $T \geq 1$ number of iterations,
$\eta > 0$ learning rate, $N \geq 1$ number of Monte-Carlo particles.
Output $\theta^{(0)}, \ldots, \theta^{(T-1)}$
for $t = 0 \ldots T-1$ **do**
  Sample $i(t) \sim \text{Unif}\{1, \ldots, n\}$
  Sample $\mathbf{V}_k \sim \pi^{(t)} (1 \leq k \leq N)$
  Compute self-normalized gradient $\widehat{g}^{(t)}$
  Update $\theta^{(t+1)} = \theta^{(t)} + \eta \widehat{g}^{(t)}$
end

Convergence guarantees of SGD with biased gradients

Theorem ([Ajalloeian and Stich, 2021])

Let $\epsilon > 0$ and assume ℓ is $L-$smooth ($\nabla^2 \ell$ bounded by L). If for all $t \geq 1$

$$MSE(\widehat{g}^{(t)}) = \mathbb{E}\left[\left\|\widehat{g}^{(t)} - \nabla_\theta \ell\left(\theta^{(t)}\right)\right\|^2\right] < \infty,$$

$T \geq \frac{1}{\epsilon^2 + \xi^2}$ and η is chosen wisely, then the sequence $\left(\theta^{(t)}\right)_{0 \leq t \leq T-1}$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla_\theta \ell\left(\theta^{(t)}\right)\right\|^2\right] \leq K\left(\epsilon + \xi\right),$$

with ξ is a constant growing with the bias of the estimator.

Bias control

Theorem ([Agapiou et al., 2017])

For all $t \geq 0$ *and* $1 \leq i \leq n$*, if* $\mathbb{E}_{\pi(\cdot;\theta^{(t)},i)}\left[\|\nabla_\theta \log\left(p_{\theta^{(t)}}(\mathbf{Y}_i,\mathbf{V})\right)\|_1^4\right]$ *is finite and the weights* $\frac{p_\theta(\mathbf{Y}_i,\cdot)}{\pi(\cdot;\theta^{(t)},i)}$ *are bounded almost surely, we have*

$$MSE(\widehat{g}^{(t)}) = \mathbb{E}\left[\left\|\widehat{g}^{(t)} - \nabla_\theta \ell\left(\theta^{(t)}\right)\right\|^2\right] = o\left(\frac{1}{N}\right) \tag{1}$$

and

$$B(\widehat{g}^{(t)}) = \left\|\mathbb{E}\left[\widehat{g}^{(t)}\right] - \nabla_\theta \ell\left(\theta^{(t)}\right)\right\|^2 = o\left(\frac{1}{N}\right). \tag{2}$$

Recap

$$\frac{p_\theta(\boldsymbol{Y}_i,\cdot)}{\pi(\cdot)} \overset{a.s.}{<} \infty \quad + \quad \nabla \log p_\theta(\boldsymbol{Y}_i,\cdot) \in L^4(\pi)$$

$\Downarrow$ $\qquad\qquad$ $\Downarrow$

L-smoothness $\quad + \quad$ $MSE(\widehat{g}^{(t)}) < \infty$

$$\Downarrow$$

Convergence à $\epsilon +$ biais près

Choice of $\pi(\cdot;\theta,i)$

- For a proposal g, a reasonable estimate is reached [Chatterjee and Diaconis, 2018] once

$$N \approx e^{KL(p_\theta(\cdot,\mathbf{Y}_i)||g)}$$

so that we wish to take

$$\pi^\star(\cdot;\theta,i) = \underset{g\in\mathcal{F}}{\operatorname{argmin}}\, KL(p_\theta(\cdot\mid\mathbf{Y}_i)\|g)$$

- Here we take:

$$\mathcal{F} = \left\{\mathcal{N}_q(\mathbf{m},\mathbf{S}) \mid \mathbf{m}\in\mathbb{R}^q,\, \mathbf{S}\in\mathcal{S}_q^+\right\}.$$

- After a few computations, we get

$$\pi^\star(\cdot;\theta,i) = \mathcal{N}_q\left(\mathbb{E}\left[\boldsymbol{W}|\boldsymbol{Y}_i\right],\mathbb{V}\left[\boldsymbol{W}|\boldsymbol{Y}_i\right]\right)$$

Integrability and boundedness

- For the Poisson and Binomial case, the integrability condition is ensured.
- The weights are bounded if $\pi^\star(\cdot) \geq K \exp\left(-\frac{\|\cdot\|^2}{2}\right)$, which cannot be ensured.
- Solution to ensure boundedness $\Longrightarrow$ mix $\pi^\star$ with a "defensive" proposal with higher variance:

$$\pi^\star_\alpha(\cdot, \theta, i) = (1-\alpha)\,\pi^\star(\cdot, \theta, i) + \alpha \mathcal{N}(\mathbb{E}\left[\boldsymbol{W} | \boldsymbol{Y}_i\right], \delta \boldsymbol{I}_q)$$

with $\delta > 1$ and $0 < \alpha < 1$.

Recap

$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$

$\pi(\boldsymbol{V}) \leq K \exp\left(-\frac{\|\boldsymbol{V}\|^2}{2}\right)$

$\Downarrow$

$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$

Binomial ou Poisson

$\Downarrow$

$\frac{p_\theta(\boldsymbol{Y}_i, \cdot)}{\pi(\cdot)} \overset{a.s.}{<} \infty \quad + \quad \nabla \log p_\theta(\boldsymbol{Y}_i, \cdot) \in L^4(\pi)$

$\Downarrow \qquad \Downarrow$

L-smoothness $+$ $MSE(\widehat{g}^{(t)}) < \infty$

$\Downarrow$

Convergence à $\epsilon +$ biais près

L-smoothness

- It cannot be shown that $\theta \mapsto \ell(\theta)$ is L-smooth.
- Moreover, the learning rate η must be set as a function of the supremum of the bias $\implies$ All the bounds must be uniform on θ.
- Solution $\implies$ restrict ourselves to $\theta \in \mathcal{X}$ with $\mathcal{X}$ a compact convex subset.
- Need to adapt SGD to Projected SGD and the convergence proof.
- [Mai and Johansson, 2021] proves convergence for projected SGD in a non-convex setting $\implies$ only the bias must be added.

Conclusion

- Scaling well with the number of samples n thanks to SGD.
- Relatively high number of dimensions can be selected thanks to low-dimensional sampling.
- Still need to adapt to projected gradient to get theoretical guarantees.

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017).
Importance sampling: Intrinsic dimension and computational cost.

Ajalloeian, A. and Stich, S. U. (2021).
On the convergence of sgd with biased gradients.

Chatterjee, S. and Diaconis, P. (2018).
The sample size required in importance sampling.
The Annals of Applied Probability, 28(2):1099–1135.

Mai, V. V. and Johansson, M. (2021).
Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization.