Stochastic Variable Metric Forward-Backward with variance reduction

Gersende Fort
CNRS
Institut de Mathématiques de Toulouse, France

LOL 2022 - Learning and Optimization in Luminy, CIRM, October 2022.

In collaboration with

- Eric Moulines, Ecole Polytechnique, CMAP, France

Talk based on the paper:

- *Stochastic Variable Metric Proximal Gradient with variance reduction for non-convex composite optimization*
  by G. Fort and E. Moulines.
  HAL-03781216

Partly funded by
Fondation Simone et Cino Del Duca, Project OpSiMorE

I. Problem and Motivations

Stochastic Optimization

- Solve

$$\boxed{0 \in \frac{1}{n}\sum_{i=1}^{n} G_i(s) + \partial g(s) \qquad\qquad s \in \mathbb{R}^q}$$

where
- the fct $g : \mathbb{R}^q \to (-\infty, +\infty]$ is lower semi-continuous, convex
- with domain $\mathcal{S} := \{s \in \mathbb{R}^q : g(s) < +\infty\}$
- the function $G_i : \mathbb{R}^q \to \mathbb{R}^q$

- **Requirements:** Design and study an algorithm such that
  - possibly Preconditioned operators

$$B^{-1}\, G_i(s) \qquad B \text{ is a } q \times q \text{ positive definite matrix}$$

  - possibly approximated preconditioned operators
  - finite-sum challenge: solution via a stochastic procedure with variance reduction.

Appli. 1: Gradient-based algorithms (1/2)

$$\mathrm{argmin}_{s \in \mathbb{R}^q} \qquad \frac{1}{n}\sum_{i=1}^{n} \ell_i(s) + g(s)$$

- Ex. in statistical Learning
  g is a regularization term, or an a priori on the parameter s
  ℓ_i is a loss function associated to the example $\#i$

- When
  - The fct $g : \mathbb{R}^q \to (-\infty, +\infty]$ is lower semi-continuous, **convex**
  - with domain $\mathcal{S} := \{s \in \mathbb{R}^q : g(s) < +\infty\}$
  - $s \mapsto \ell_i(s)$ is C^1 on $\mathcal{S}$ **no convexity assumptions** on the ℓ_i's

Appli. 1: Gradient-based algorithms (1/2)

$$\mathrm{argmin}_{s \in \mathbb{R}^q} \qquad \frac{1}{n}\sum_{i=1}^{n} \ell_i(s) + g(s)$$

- Ex. in statistical Learning
  g is a regularization term, or an a priori on the parameter s
  ℓ_i is a loss function associated to the example $\#i$

- When
  - The fct $g : \mathbb{R}^q \to (-\infty, +\infty]$ is lower semi-continuous, **convex**
  - with domain $\mathcal{S} := \{s \in \mathbb{R}^q : g(s) < +\infty\}$
  - $s \mapsto \ell_i(s)$ is C^1 on $\mathcal{S}$ **no convexity assumptions** on the ℓ_i's

- Often, "solved" by

$$0 \in \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(s) + \partial g(s) \qquad\qquad s \in \mathbb{R}^q$$

Appli. 1: Gradient-based algorithms (2/2)

$$0 \in \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(s) + +\partial g(s) \qquad\qquad s \in \mathbb{R}^q$$

Under smoothness assumptions on the ℓ_i's,

- Forward-Backward splitting:

$$s_{t+\frac{1}{2}} = s_t - \gamma_{t+1}\ \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(s_t)$$

$$s_{t+1} = \mathrm{prox}_{\gamma_{t+1}\, g}\left(s_{t+\frac{1}{2}}\right)$$

where Moreau (1965)

$$\mathrm{prox}_{\gamma\, g}(s) := \mathrm{argmin}_{\mathbb{R}^q} \gamma\, g(\cdot) + \tfrac{1}{2}\| \cdot - s\|^2$$

Appli. 1: Gradient-based algorithms (2/2)

$$0 \in \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(s) + +\partial g(s) \qquad\qquad s \in \mathbb{R}^q$$

Under smoothness assumptions on the ℓ_i's,

- Forward-Backward splitting:

$$s_{t+\frac{1}{2}} = s_t - \gamma_{t+1} \ \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \nabla \ell_i(s_t)$$

$$s_{t+1} = \mathrm{prox}_{\gamma_{t+1}\, g}\left(s_{t+\frac{1}{2}}\right)$$

where Moreau (1965)

$$\mathrm{prox}_{\gamma\, g}(s) := \mathrm{argmin}_{\mathbb{R}^q} \gamma\, g(\cdot) + \tfrac{1}{2}\|\cdot - s\|^2$$

Appli. 1: Gradient-based algorithms (2/2)

$$0 \in \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(s) + +\partial g(s) \qquad\qquad s \in \mathbb{R}^q$$

Under smoothness assumptions on the ℓ_i's,

- Forward-Backward splitting:

$$s_{t+\frac{1}{2}} = s_t - \gamma_{t+1}\ \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} B^{-1}\, \nabla \ell_i(s_t)$$

$$s_{t+1} = \mathrm{prox}^{B}_{\gamma_{t+1}\, g}\left(s_{t+\frac{1}{2}}\right)$$

where see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$\mathrm{prox}^{B}_{\gamma\, g}(s) := \mathrm{argmin}_{\mathbb{R}^q} \quad \gamma\, g(\cdot) + \tfrac{1}{2}\| \cdot - s\|^2_B$$

- Remark: Preconditioned gradients

- for acceleration Chouzenoux et al (2014), Repetti et al (2014)
- variable metric on the gradient $\Longrightarrow$ variable metric on the proximal

Appli. 1: Gradient-based algorithms (2/2)

$$0 \in \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(s) + +\partial g(s) \qquad\qquad s \in \mathbb{R}^q$$

Under smoothness assumptions on the ℓ_i's,

- Forward-Backward splitting:

$$s_{t+\frac{1}{2}} = s_t - \gamma_{t+1} \, \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} B^{-1} \, \nabla \ell_i(s_t)$$

$$s_{t+1} = \mathrm{prox}^{B}_{\gamma_{t+1}\, g}\left(s_{t+\frac{1}{2}}\right)$$

where see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$\mathrm{prox}^{B}_{\gamma\, g}(s) := \mathrm{argmin}_{\mathbb{R}^q} \quad \gamma\, g(\cdot) + \frac{1}{2}\| \cdot - s\|^2_B$$

Let us go beyond:

- variance reduction of the mini batch approximation

- approximated gradient: $\widehat{\nabla \ell_i(s_t)}$

- Remark: Preconditioned gradients

- for acceleration Chouzenoux et al (2014), Repetti et al (2014)
- variable metric on the gradient $\Longrightarrow$ variable metric on the proximal

Appli. 2: Expectation-Maximization for curved exponential families (1/2)

Dempster et al (1977), Wu (1983)

- For inference by ML in latent variable models ex. mixture models

$$\mathrm{argmin}_{\theta \in \Theta} - \frac{1}{n} \sum_{i=1}^{n} \log \int_{\mathsf{Z}} p(Y_i, z; \theta) \, \mathsf{d}\mu(z) \qquad \Theta \subseteq \mathbb{R}^d$$

with complete data likelihood from the curved exponential family McLachlan and Krishnan (2008)

$$\log p(Y_i, z; \theta) = \langle s_i(z), \phi(\theta) \rangle - \psi(\theta)$$

Appli. 2: Expectation-Maximization for curved exponential families (1/2)

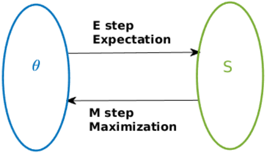
Dempster et al (1977), Wu (1983)

- For inference by ML in latent variable models ex. mixture models

$$\mathrm{argmin}_{\theta \in \Theta} - \frac{1}{n} \sum_{i=1}^{n} \log \int_{\mathsf{Z}} p(Y_i, z; \theta) \, \mathsf{d}\mu(z) \qquad \Theta \subseteq \mathbb{R}^d$$

with complete data likelihood from the curved exponential family McLachlan and Krishnan (2008)

$$\log p(Y_i, z; \theta) = \langle s_i(z), \phi(\theta) \rangle - \psi(\theta)$$



M step $\theta := \mathsf{T}(s)$

E step $s := n^{-1} \sum_{i=1}^{n} \bar{s}_i(\theta)$

$$\bar{s}_i(\theta) := \mathbb{E}\left[s_i(Z); \theta, i\right]$$

Appli. 2: Expectation-Maximization for curved exponential families (1/2)

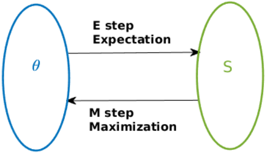
Dempster et al (1977), Wu (1983)

- For inference by ML in latent variable models ex. mixture models

$$\mathrm{argmin}_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^{n}\log\int_{\mathsf{Z}} p(Y_i,z;\theta)\,\mathsf{d}\mu(z) \qquad \Theta\subseteq\mathbb{R}^d$$

with complete data likelihood from the curved exponential family McLachlan and Krishnan (2008)

$$\log p(Y_i,z;\theta) = \langle s_i(z), \phi(\theta)\rangle - \psi(\theta)$$



M step $\theta := \mathsf{T}(s)$

E step $s := n^{-1}\sum_{i=1}^{n}\bar{s}_i(\theta)$

$$\bar{s}_i(\theta) := \mathbb{E}\left[s_i(Z);\theta,i\right]$$

- In the θ space

$$\theta_{t+1} = \mathsf{T}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{s}_i(\theta_t)\right) \qquad \mathsf{T}(n^{-1}\sum_{i=1}^{n}\bar{s}_i(\theta)) - \theta = 0$$

- In the statistic space

$$s_{t+1} = \frac{1}{n}\sum_{i=1}^{n}\bar{s}_i\left(\mathsf{T}\left(s_t\right)\right) \qquad \frac{1}{n}\sum_{i=1}^{n}\bar{s}_i(\mathsf{T}(s)) - s = 0$$

Appli. 2: Expectation-Maximization for curved exponential families (2/2)

- EM in the statistic space solves the problem

$$\boxed{0 = \frac{1}{n} \sum_{i=1}^{n} \left\{ \bar{s}_i \left(\mathsf{T}(s) \right) - s \right\} \quad \textbf{and} \quad s \in \mathcal{S}} \qquad ,$$

Appli. 2: Expectation-Maximization for curved exponential families (2/2)

- EM in the statistic space solves the problem

$$\boxed{0 = \frac{1}{n}\sum_{i=1}^{n}\{\bar{s}_i\left(\mathsf{T}(s)\right) - s\} \quad \textbf{and} \quad s \in \mathcal{S}},$$

- EM in the statistic space is a preconditioned gradient algorithm Delyon et al (1999), Fort et al (2020)

$$\frac{1}{n}\sum_{i=1}^{n}\bar{s}_i\left(\mathsf{T}(s)\right) - s = -\mathsf{B}(s)^{-1}\,\nabla W(s)$$

Appli. 2: Expectation-Maximization for curved exponential families (2/2)

- EM in the statistic space solves the problem

$$\boxed{0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\bar{s}_i\left(\mathsf{T}(s)\right) - s\right\} \quad \textbf{and} \quad s \in \mathcal{S}},$$

- EM in the statistic space is a preconditioned gradient algorithm Delyon et al (1999), Fort et al (2020)

$$\frac{1}{n}\sum_{i=1}^{n}\bar{s}_i\left(\mathsf{T}(s)\right) - s = -\mathsf{B}(s)^{-1}\,\nabla W(s)$$

- Inexact preconditioned gradients: Celeux and Diebolt (1985), Wei and Tanner (1990), Delyon et al (1999), Fort and Moulines (2003)

$$\bar{s}_i(\tau) := \int_{\mathsf{Z}} s_i(z)\,\frac{p(Y_i, z; \tau)\,\mathsf{d}\mu(z)}{\int p(Y_i, u; \tau)\,\mathsf{d}\mu(u)} \qquad \text{random approximations, MCMC}$$

Appli. 2: Expectation-Maximization for curved exponential families (2/2)

- EM in the statistic space solves the problem

$$\boxed{0 = \frac{1}{n}\sum_{i=1}^{n} \left\{ \bar{s}_i \left(\mathsf{T}(s) \right) - s \right\} \quad \textbf{and} \quad s \in \mathcal{S}} \quad ,$$

- EM in the statistic space is a preconditioned gradient algorithm Delyon et al (1999), Fort et al (2020)

$$\frac{1}{n}\sum_{i=1}^{n} \bar{s}_i \left(\mathsf{T}(s) \right) - s = -\mathsf{B}(s)^{-1} \nabla W(s)$$

- Inexact preconditioned gradients: Celeux and Diebolt (1985), Wei and Tanner (1990), Delyon et al (1999), Fort and Moulines (2003)

$$\bar{s}_i(\tau) := \int_{\mathsf{Z}} s_i(z) \frac{p(Y_i, z; \tau)\, \mathrm{d}\mu(z)}{\int p(Y_i, u; \tau)\, \mathrm{d}\mu(u)} \qquad \text{random approximations, MCMC}$$

- Incremental EM algorithms: the finite sum setting addressed via stochastic EM in the statistic space. Neal and Hinton (1998), Ng and McLachlan (2003), Cappé and Moulines (2009), Chen et al (2018), Karimi et al (2019), Fort et al (2020, 2021)

II. Contributions

Contributions

$$\boxed{0 \in \frac{1}{n} \sum_{i=1}^{n} G_i(s) + \partial g(s) \qquad s \in \mathbb{R}^q}$$

- see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$s = \mathrm{prox}_{\gamma\, g}^{B}(s + \gamma\ h) \qquad \text{iff} \quad 0 \in -Bh + \partial g(s)$$

$\hookrightarrow$ (Variable Metric) Forward-Backward

Contributions

$$\boxed{0 \in \frac{1}{n} \sum_{i=1}^{n} G_i(s) + \partial g(s) \qquad s \in \mathbb{R}^q}$$

- see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$s = \mathrm{prox}_{\gamma\, g}^{B}(s + \gamma\ h) \qquad \text{iff} \quad 0 \in -Bh + \partial g(s)$$

$\hookrightarrow$ (Variable Metric) Forward-Backward

- We propose an algorithm

- forward step: preconditioned forward operators

$$\mathsf{h}_i(s, B) := -B^{-1}\, G_i(s)$$

possibly approximated $\widehat{\mathsf{h}_i(s, B)}$,
- addresses the finite sum setting by minibatches & variance reduction
- backward step: proximity operator associated to g $\mathrm{prox}_{\gamma g}$, assumed exact

Contributions

$$\boxed{0 \in \frac{1}{n}\sum_{i=1}^{n} G_i(s) + \partial g(s) \qquad\qquad s \in \mathbb{R}^q}$$

- see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$s = \mathrm{prox}_{\gamma\, g}^{B}(s + \gamma\ h) \qquad \text{iff} \quad 0 \in -Bh + \partial g(s)$$

$\hookrightarrow$ (Variable Metric) Forward-Backward

- We propose an algorithm

- forward step: preconditioned forward operators

$$\mathsf{h}_i(s, B) := -B^{-1}\, G_i(s)$$

possibly approximated $\widehat{\mathsf{h}_i(s, B)}$,
- addresses the finite sum setting by minibatches & variance reduction
- backward step: proximity operator associated to g $\quad \mathrm{prox}_{\gamma g}$, assumed exact

- We provide explicit convergence bounds in expectation

discuss the complexity of the algorithm (w.r.t. n and the tolerance ϵ)
discuss the impact of the approximations on the h_i's
in the **non convex** case.

For gradient-based algorithms

g	non-cvx	finite sum	red var	Precond	Approx. forward h_i's	refs
	✓	✓	✓			Ghadimi and Lan (2013), Reddi et al (2016) Allen-Zhu and Hazan (2016) Nguen et al (2017), Allen-Zhu (2018) Fang et al (2018), Dongruo et al (2020)
✓	✓	✓				Ghadimi et al (2016), Karimi et al (2016)
✓	✓	✓	✓			Li and Li (2018), Wang et al (2019) Zhang and Xiao (2019), Nhan et al (2020) Metel and Takeda (2021)
✓	✓			✓	✓ unbiased & bounded	Yun et al (2021)
✓		✓			✓ (un)biased	Atchade et al (2017)
✓	✓	✓	✓	✓	✓ (un)biased	our contribution

For EM algorithms

g	non-cvx	finite sum	red var	Precond	Approx. forward h_i's	refs
	✓			✓	✓ (un)biased	Celeux and Doebolt (1985) Wei and Tanner (1990) Delyon et al (1999) Fort and Moulines (2003)
	✓	✓		✓		Neal and Hinton (1998) Cappé and Moulines (2009)
	✓	✓	✓	✓		Chen et al (2018), Karimi et al (2019) Fort et al (2020)
✓	✓	✓	✓	✓	✓ (un)biased	our contribution

III. The 3P-SPIDER algorithm

Perturbed Proximal Preconditioned Stochastic Path-Integrated Differential EstimatoR

3P-SPIDER

Algorithm: 3P-SPIDER

$\hat{S}_{0,k_0^{\text{in}}} = \hat{S}_{\text{init}}, \qquad B_{0,k_0^{\text{in}}} = B_{\text{init}}$

for $\underline{t = 1, \cdots, k^{\text{out}}}$ **do**

$\hat{S}_{t,0} = \hat{S}_{t-1,k_{t-1}^{\text{in}}}, \quad \hat{S}_{t,-1} = \hat{S}_{t-1,k_{t-1}^{\text{in}}},$

$B_{t,0} = B_{t-1,k_{t-1}^{\text{in}}}$

Sample a batch $\mathcal{B}_{t,0}$ of size b'_t in $\{1, \cdots, n\}$, with or without replacement.

For all $i \in \mathcal{B}_{t,0}$, compute $\delta_{t,0,i}$ equal to or approximating $\mathsf{h}_i(\hat{S}_{t,0}, B_{t,0})$.

$\mathsf{S}_{t,0} = (\mathsf{b}'_t)^{-1} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i}$

for $\underline{k = 0, \cdots, k_t^{\text{in}} - 1}$ **do**

Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b in $\{1, \cdots, n\}$, with or without replacement

Choose $B_{t,k+1}$, a positive definite matrix

For all $i \in \mathcal{B}_{t,k+1}$, compute $\delta_{t,k+1,i} \approx \mathsf{h}_i(\hat{S}_{t,k}, B_{t,k+1}) - \mathsf{h}_i(\hat{S}_{t,k-1}, B_{t,k})$

$\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \delta_{t,k+1,i}$

$\hat{S}_{t,k+\frac{1}{2}} = \hat{S}_{t,k} + \gamma_{t,k+1} \, \mathsf{S}_{t,k+1}$

$\hat{S}_{t,k+1} = \text{prox}_{t,k}(\hat{S}_{t,k+\frac{1}{2}}),$ where $\text{prox}_{t,k} := \text{prox}_{\gamma_{t,k+1}\, g}^{B_{t,k+1}}.$

Algorithm: Variable Metric Forward-Backward + Finite sum (with variance reduction)

$\hat{S}_{0,k_0^{\text{in}}} = \hat{S}_{\text{init}}, \qquad B_{0,k_0^{\text{in}}} = B_{\text{init}}$

for $t = 1, \cdots, k^{\text{out}}$ **do**

$\hat{S}_{t,0} = \hat{S}_{t-1,k_{t-1}^{\text{in}}}, \quad \hat{S}_{t,-1} = \hat{S}_{t-1,k_{t-1}^{\text{in}}},$

$B_{t,0} = B_{t-1,k_{t-1}^{\text{in}}}$

Sample a batch $\mathcal{B}_{t,0}$ of size b'_t in $\{1, \cdots, n\}$, with or without replacement.

For all $i \in \mathcal{B}_{t,0}$, compute $\delta_{t,0,i}$ equal to or approximating $\mathsf{h}_i(\hat{S}_{t,0}, B_{t,0})$.

$\mathsf{S}_{t,0} = (\mathsf{b}'_t)^{-1} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i}$

for $k = 0, \cdots, k_t^{\text{in}} - 1$ **do**

Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b in $\{1, \cdots, n\}$, with or without replacement

Choose $B_{t,k+1}$, a positive definite matrix

$\mathsf{S}_{t,k+1} = \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \mathsf{h}_i(\hat{S}_{t,k}, B_{t,k+1})$

$\hat{S}_{t,k+\frac{1}{2}} = \hat{S}_{t,k} + \gamma_{t,k+1} \, \mathsf{S}_{t,k+1}$

$\hat{S}_{t,k+1} = \text{prox}_{t,k}(\hat{S}_{t,k+\frac{1}{2}}),$ where $\text{prox}_{t,k} := \text{prox}_{\gamma_{t,k+1}\, g}^{B_{t,k+1}}$.

Algorithm: Variable Metric Forward-Backward + Finite sum (with variance reduction)

$\hat{S}_{0,k_0^{\text{in}}} = \hat{S}_{\text{init}}, \qquad B_{0,k_0^{\text{in}}} = B_{\text{init}}$

for $t = 1, \cdots, k^{\text{out}}$ **do**

$\hat{S}_{t,0} = \hat{S}_{t-1,k_{t-1}^{\text{in}}}, \quad \hat{S}_{t,-1} = \hat{S}_{t-1,k_{t-1}^{\text{in}}},$

$B_{t,0} = B_{t-1,k_{t-1}^{\text{in}}}$

Sample a batch $\mathcal{B}_{t,0}$ of size b'_t in $\{1, \cdots, n\}$, with or without replacement.

For all $i \in \mathcal{B}_{t,0}$, compute $\delta_{t,0,i}$ equal to or approximating $\mathsf{h}_i(\hat{S}_{t,0}, B_{t,0})$.

$\mathsf{S}_{t,0} = (\mathsf{b}'_t)^{-1} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i}$

for $k = 0, \cdots, k_t^{\text{in}} - 1$ **do**

Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b in $\{1, \cdots, n\}$, with or without replacement

Choose $B_{t,k+1}$, a positive definite matrix

$\mathsf{S}_{t,k+1} = \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \mathsf{h}_i(\hat{S}_{t,k}, B_{t,k+1}) + \left(\mathsf{S}_{t,k} - \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t+1}} \mathsf{h}_i(\hat{S}_{t,k-1}, B_{t,k}) \right)$

$\hat{S}_{t,k+\frac{1}{2}} = \hat{S}_{t,k} + \gamma_{t,k+1} \, \mathsf{S}_{t,k+1}$

$\hat{S}_{t,k+1} = \mathrm{prox}_{t,k}(\hat{S}_{t,k+\frac{1}{2}}),$ where $\mathrm{prox}_{t,k} := \mathrm{prox}_{\gamma_{t,k+1}\, g}^{B_{t,k+1}}.$

Zoom on the variance reduction by SPIDER Fang et al (2018), Nguyen et al (2017), Wang et al (2019)

- The SPIDER control variate: if S_t approximates $n^{-1}\sum_{i=1}^n \mathsf{p}_i(s_{t-1})$ then

$$\mathsf{S}_{t+1} := \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \mathsf{p}_i(s_t) + \left(\mathsf{S}_t - \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \mathsf{p}_i(s_{t-1}) \right) \qquad \approx \frac{1}{n}\sum_{i=1}^n \mathsf{p}_i(s_t)$$

- Biased control variate ! F. and Moulines (2022, Proposition 7.3.)

$$\mathbb{E}\left[\mathsf{S}_{t+1} \middle| \mathsf{Past}_t \right] \neq \frac{1}{n}\sum_{i=1}^n \mathsf{p}_i(s_t)$$

Refresh regularly the control variate:

Outer loops

- initialize the control variate
- repeat k^{in} *inner loops* of the stochastic VMFB algorithm

Algorithm: 3P-SPIDER = VMFB + Finite sum with var red + Perturbed forward step

$\hat{S}_{0,k_0^{\text{in}}} = \hat{S}_{\text{init}}, \qquad B_{0,k_0^{\text{in}}} = B_{\text{init}}$

for $t = 1, \cdots, k^{\text{out}}$ **do**

$\hat{S}_{t,0} = \hat{S}_{t-1,k_{t-1}^{\text{in}}}, \quad \hat{S}_{t,-1} = \hat{S}_{t-1,k_{t-1}^{\text{in}}} \quad B_{t,0} = B_{t-1,k_{t-1}^{\text{in}}}$

Sample a batch $\mathcal{B}_{t,0}$ of size b'_t in $\{1, \cdots, n\}$, with or without replacement.

For all $i \in \mathcal{B}_{t,0}$, compute $\delta_{t,0,i}$ equal to or approximating $\mathsf{h}_i(\hat{S}_{t,0}, B_{t,0})$.

$\mathsf{S}_{t,0} = (\mathsf{b}'_t)^{-1} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i}$

for $k = 0, \cdots, k_t^{\text{in}} - 1$ **do**

Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b in $\{1, \cdots, n\}$, with or without replacement

Choose $B_{t,k+1}$, a positive definite matrix

For all $i \in \mathcal{B}_{t,k+1}$, compute $\delta_{t,k+1,i} \approx \mathsf{h}_i(\hat{S}_{t,k}, B_{t,k+1}) - \mathsf{h}_i(\hat{S}_{t,k-1}, B_{t,k})$

$\mathsf{S}_{t,k+1} = \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \delta_{t,k+1,i} + \mathsf{S}_{t,k}$

$\hat{S}_{t,k+\frac{1}{2}} = \hat{S}_{t,k} + \gamma_{t,k+1} \, \mathsf{S}_{t,k+1}$

$\hat{S}_{t,k+1} = \operatorname{prox}_{t,k}(\hat{S}_{t,k+\frac{1}{2}}), \qquad$ where $\operatorname{prox}_{t,k} := \operatorname{prox}_{\gamma_{t,k+1}\, g}^{B_{t,k+1}}$.

IV. On an example

Logistic regression with random effects: the model Details in F.and Moulines (2022)

- Given :

- Observations $Y_1, \cdots, Y_n$ in $\{-1, 1\}$; independent
- Covariates $X_1, \cdots, X_n$ in $\mathbb{R}^d$

- Random effects Z_i

$$\mathbb{P}(Y_i = 1 \Big| Z_i) = \frac{1}{1 + \exp(X_i^\top Z_i)} \qquad Z_i \overset{i.i.d.}{\sim} \mathcal{N}_d(\theta, \sigma^2 \mathrm{I}).$$

Logistic regression with random effects: the model Details in F.and Moulines (2022)

- Given :

- Observations $Y_1, \cdots, Y_n$ in $\{-1, 1\}$; independent
- Covariates $X_1, \cdots, X_n$ in $\mathbb{R}^d$

- Random effects Z_i

$$\mathbb{P}(Y_i = 1 \big| Z_i) = \frac{1}{1 + \exp(X_i^\top Z_i)} \qquad Z_i \overset{i.i.d.}{\sim} \mathcal{N}_d(\theta, \sigma^2 \mathrm{I}).$$

- Estimation of θ by penalized ML

$$\mathrm{argmin}_{\theta \in \mathbb{R}^d} \; -\frac{1}{n} \sum_{i=1}^{n} \log \int \frac{1}{1 + \exp(Y_i \, X_i^\top z_i)} \exp\left(-\frac{1}{2\sigma^2} \|\theta - z_i\|^2\right) \, \mathrm{d}z_i + \tau \|\theta\|^2.$$

- Remark: the minimizers are in a compact set $\mathcal{K} := \{\theta \in \mathbb{R}^d : \|\theta\|^2 \leq \ln(4)/\tau\}$.

In this example

- $n = 24\,989$ examples; $d = 21$.

- 3P-SPIDER $\equiv$ an EM in the statistic space.

- $\mathrm{prox}_{\gamma g}^{B}$: projection on a compact set.
- MCMC approximation of the h_i's

In this example

- $n = 24\,989$ examples; $d = 21$.
- 3P-SPIDER $\equiv$ an EM in the statistic space.
- $\mathrm{prox}^{B}_{\gamma g}$: projection on a compact set.
- MCMC approximation of the h_i's

	finite sum	with var red	approx h_i	updates per epoch
EM			✓	1
Online EM	✓		✓	k^{in}
3P-SPIDER	✓	✓	✓	odd: - even: k^{in}
	minibatch		MCMC	

In this example

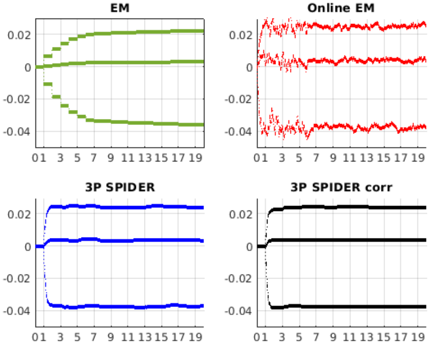
- $n = 24\,989$ examples; $d = 21$.
- 3P-SPIDER $\equiv$ an EM in the statistic space.
- $\mathrm{prox}^B_{\gamma g}$: projection on a compact set.
- MCMC approximation of the h_i's

	finite sum	with var red	approx h_i	updates per epoch
EM			✓	1
Online EM	✓		✓	k^{in}
3P-SPIDER	✓	✓	✓	odd: - even: k^{in}
	minibatch		MCMC	

- All of them, of the form $\hat{S}_{\mathrm{new}} = \mathrm{prox}^B_{\gamma g}(\hat{S}_{\mathrm{old}} + \gamma \mathcal{H})$.

  Compared through
  - the sequence $t \mapsto \theta_t$
  - a "distance" to the limiting set: $$\frac{\|\mathrm{prox}^B_{\gamma g}(\hat{S}_{\mathrm{old}} + \gamma \mathcal{H}) - \hat{S}_{\mathrm{old}}\|^2_B}{\gamma^2}$$

The variance reduction by SPIDER (1/2)



Show the benefits of

- many updates of the iterates during the first epochs (minibatch)

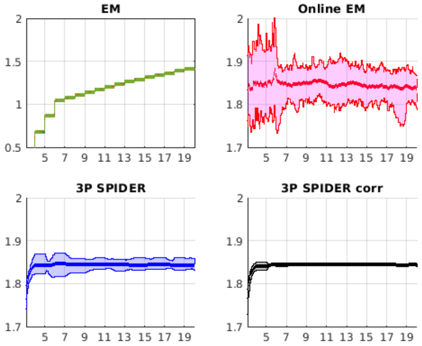
- the variance reduction to control the variability introduced by the mini-batch

- a gain when increasing the control variate effect.

Estimation of three components of θ.

Evolution of the three components of θ by Δ_r^{EM} in green (top, left), Δ_r^{OEM} in red (top, right), $\Delta_{t,k+1}$ for 3P-SPIDER in blue (bottom, left) and $\Delta_{t,k+1}$ for 3P-SPIDER corr in black (bottom, right), as a function of the number of epochs

The variance reduction by SPIDER (2/2)



Show the benefits of

- many updates of the iterates during the first epochs (minibatch)

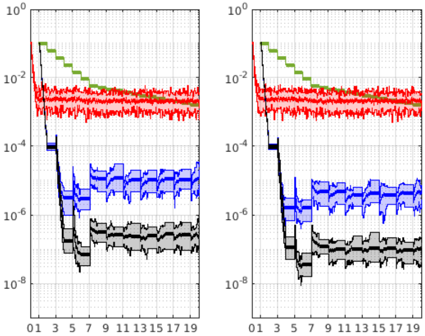
- the variance reduction to control the variability introduced by the mini-batch

- a gain when increasing the control variate effect.

Evolution of the squared norm of the iterates.
Mean value over 25 runs; (shadowed) min/max fluctuations

Evolution of $\|\hat{S}_r^{\mathtt{EM}}\|^2$ in green (top, left), $\|\hat{S}_r^{\mathtt{OEM}}\|^2$ in red (top, right), $\Delta_{t,k+1}$ for 3P-SPIDER in blue (bottom, left) and $\Delta_{t,k+1}$ for 3P-SPIDER corr in black (bottom, right), as a function of the number of epochs.

Fluctuations at convergence – Nbr of Monte Carlo points



Show the benefits of

- (same as before)

- a larger number of Monte Carlo points

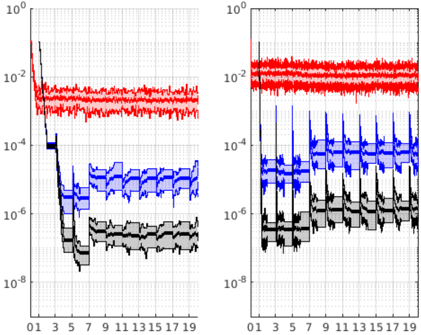
A larger step size from epoch #7

Two strategies for the number of Monte Carlo points when approximating h_i
Larger number on the right

Mean value over 25 runs; (shadowed) min/max fluctuations

Evolution of $\Delta_r^{\mathtt{EM}}$ in green, $\Delta_r^{\mathtt{OEM}}$ in red, $\Delta_{t,k+1}$ for 3P-SPIDER in blue and $\Delta_{t,k+1}$ for 3P-SPIDER corr in black, as a function of the number of epochs. [left] $m^0 = m^t = 2\lceil\sqrt{n}\rceil$, [right] $m^0 = m^t = 5\lceil\sqrt{n}\rceil$.

Fluctuations at convergence – Role of k^{in}



Show the benefits of

- (same as before)

- a larger minibatch size and a lower number of inner loops.

A larger step size from epoch $\#7$

Two strategies for the number of inner loops per epoch
Larger number on the right ($\Rightarrow$ smaller minibatch size)

Mean value over 25 runs; (shadowed) min/max fluctuations

Evolution of $\Delta_r^{\texttt{OEM}}$ in red, $\Delta_{t,k+1}$ for 3P-SPIDER in blue and $\Delta_{t,k+1}$ for 3P-SPIDER corr in black, as a function of the number of epochs. [left] $k^{\text{in}} = \lceil \sqrt{n}/10 \rceil$ and b $= \lceil n/k^{\text{in}} \rceil$. [right] $k^{\text{in}} = \lceil \sqrt{n}/2 \rceil$ and b $= \lceil n/k^{\text{in}} \rceil$.

V. Convergence Analysis

in the case $\quad B_{t,k+1} := \mathsf{B}(\hat{S}_{t,k})$

Approximate ϵ-stationary point

$$\boxed{0 \in \frac{1}{n}\sum_{i=1}^{n} G_i(s) + \partial g(s) \qquad\qquad s \in \mathbb{R}^q}$$

- For any $\gamma > 0$, B positive definite and $h \in \mathbb{R}^q$ see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$s = \mathrm{prox}_{\gamma\, g}^{B}(s + \gamma\; h) \qquad \text{iff} \quad 0 \in -Bh + \partial g(s).$$

Approximate ϵ-stationary point

$$\boxed{0 \in \frac{1}{n} \sum_{i=1}^n G_i(s) + \partial g(s) \qquad\qquad s \in \mathbb{R}^q}$$

- For any $\gamma > 0$, B positive definite and $h \in \mathbb{R}^q$ see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$s = \mathrm{prox}^B_{\gamma\, g}(s + \gamma\, h) \qquad \text{iff} \quad 0 \in -Bh + \partial g(s).$$

- A control along iterations of

$$\Delta^{\star}_{t,k+1} := \mathbb{E}\left[\frac{\|\mathrm{prox}^{\mathsf{B}(\hat{S}_{t,k})}_{\gamma_{t,k+1}\, g}\left(\hat{S}_{t,k} + \gamma_{t,k+1}\, \mathsf{B}(\hat{S}_{t,k})^{-1}\, n^{-1} \sum_{i=1}^n G_i(\hat{S}_{t,k})\right) - \hat{S}_{t,k}\|^2_{\mathsf{B}(\hat{S}_{t,k})}}{\gamma^2_{t,k+1}}\right]$$

Approximate ϵ-stationary point

$$\boxed{0 \in \frac{1}{n}\sum_{i=1}^n G_i(s) + \partial g(s) \qquad\qquad s \in \mathbb{R}^q}$$

- For any $\gamma > 0$, B positive definite and $h \in \mathbb{R}^q$ see e.g. Hiriart-Urruty and Lemaréchal (1996)

$$s = \mathrm{prox}^{B}_{\gamma\, g}(s + \gamma\ h) \qquad \text{iff} \quad 0 \in -Bh + \partial g(s).$$

- A control along iterations of

$$\Delta^{\star}_{t,k+1} := \mathbb{E}\left[\frac{\|\mathrm{prox}^{\mathsf{B}(\hat{S}_{t,k})}_{\gamma_{t,k+1}\, g}\left(\hat{S}_{t,k} + \gamma_{t,k+1}\,\mathsf{B}(\hat{S}_{t,k})^{-1}\, n^{-1}\sum_{i=1}^n G_i(\hat{S}_{t,k})\right) - \hat{S}_{t,k}\|^2_{\mathsf{B}(\hat{S}_{t,k})}}{\gamma^2_{t,k+1}}\right]$$

- Non-convex optimization: Lan (2020, Chapter 6)

$$\frac{1}{k^{\mathrm{out}}}\frac{1}{k^{\mathrm{in}}_t}\sum_{t=1}^{k^{\mathrm{out}}}\sum_{k=0}^{k^{\mathrm{in}}_t-1}\Delta^{\star}_{t,k+1} \qquad = \mathbb{E}\left[\Delta^{\star}_{\tau}\right] \qquad \text{random stopping rule, } \tau$$

Assumptions (1/2)

A1 The non-smooth convex function g
$g : \mathbb{R}^q \to (-\infty, +\infty]$ is proper, lower semicontinuous and convex.
Set $\mathcal{S} := \{s \in \mathbb{R}^q : g(s) < +\infty\}$.

A2 Precond. Forward operators are globally Lipschitz
Set $\bar{\mathsf{h}}_i := \mathsf{h}_i(\cdot, \mathsf{B}(\cdot))$.
For all $i \in \{1, \cdots, n\}$, $\exists\, L_i > 0$ s.t. $\forall s, s' \in \mathcal{S}$,

$$\|\bar{\mathsf{h}}_i(s) - \bar{\mathsf{h}}_i(s')\| \leq L_i \, \|s - s'\|.$$

Set $L^2 := n^{-1} \sum_{i=1}^n L_i^2$.

A3 Smooth Lyapunov function
There exists $W : \mathbb{R}^q \to \mathbb{R}$, C^1 on $\mathcal{S}$; ∇W is globally $L_{\dot{W}}$-Lipschitz on $\mathcal{S}$ s.t.

$$\forall s \in \mathcal{S}, \qquad \nabla W(s) = \frac{1}{n} \sum_{i=1}^n G_i(s)\,;$$

A3' Uniformly bounded spectrum of the preconditioning matrices
There exist positive definite matrices $\mathsf{B}(s)$ s.t. $\bar{\mathsf{h}}_i(s) = -\mathsf{B}(s)^{-1} \, G_i(s)$.
There exist $0 < v_{\min} \leq v_{\max} < +\infty$ s.t. $s \in \mathcal{S}$, $v_{\min} \| \cdot \|^2 \leq \| \cdot \|^2_{\mathsf{B}(s)} \leq v_{\max} \| \cdot \|^2$.

Assumptions (2/2)

A4 On the approximations of the h_i

- Conditionally to the past, $\{\delta_{t,k+1,i}, i \in \mathcal{B}_{t,k+1}\}$ independent.
- There exist non negative constants C_b, C_v, C_{vb} and non decreasing deterministic sequence $\{m_{t,k}, k \geq 1\}$ and $\{M_{t,k}, k \geq 1\}$ s.t. almost-surely

$$\|\frac{1}{n}\sum_{i=1}^{n} \mu_{t,k+1,i}\| \leq \frac{C_b}{m_{t,k+1}} \, .$$

$$\frac{1}{n}\sum_{i=1}^{n} \sigma^2_{t,k+1,i} \leq \frac{C_v}{M_{t,k+1}} \, ,$$

$$\frac{1}{n}\sum_{i=1}^{n} \|\mu_{t,k+1,i} - \frac{1}{n}\sum_{j=1}^{n} \mu_{t,k+1,j}\|^2 \leq \frac{C^2_{vb}}{\bar{M}^2_{t,k+1}} \, .$$

	C_b	C_v	C_{vb}
exact	0	0	0
deterministic	≥ 0	0	≥ 0
random, unbiased	0	≥ 0	0
random, biased	> 0	≥ 0	≥ 0

where

$$\text{(error)} \qquad \xi_{t,k+1,i} := \delta_{t,k+1,i} - \{\bar{\mathsf{h}}_i(\hat{S}_{t,k}) - \bar{\mathsf{h}}_i(\hat{S}_{t,k-1})\}$$

$$\text{(bias)} \qquad \mu_{t,k+1,i} := \mathbb{E}\left[\xi_{t,k+1,i} \middle| \mathrm{Past}\right]$$

$$\text{(variance)} \qquad \sigma^2_{t,k+1,i} := \mathbb{E}\left[\|\xi_{t,k+1,i} - \mu_{t,k+1,i}\|^2 \middle| \mathrm{Past}\right] \, .$$

Theorem F. and Moulines (2022, Theorem 4.1)

Assume **A1** to **A4**. Choose the step sizes $\{\gamma_{t,k+1}\}$ s.t.

$$\gamma_{t,k+1}\left(1+\frac{2C_b}{m_{t,k+1}}\right) \leq \gamma_{t,k} \;,$$

$$\Lambda_{t,k+1} := \frac{\gamma_{t,k} L_{\dot{W}}}{v_{\min}} + \gamma_{t,k}^2 L^2 \frac{2 v_{\max} k_t^{\mathrm{in}}}{v_{\min} \mathsf{b}} \left(1+\frac{2\, C_{vb}}{\sqrt{\mathsf{b}}\, \bar{M}_{t,k+1}}\right) \in (0, 1/2) \;.$$

$$\sum_{t=1}^{k^{\mathrm{out}}} \sum_{k=1}^{k_t^{\mathrm{in}}} \gamma_{t,k} \left(\frac{1}{2} - \Lambda_{t,k+1}\right) \left\{\mathbb{E}\left[\Delta_{t,k}^{\star}\right] + \mathbb{E}\left[\mathcal{D}_{t,k}^{\star}\right]\right\}$$

$$\leq \mathbb{E}\left[W(\hat{S}_{1,0}) + g(\hat{S}_{1,0})\right] - \min_{\mathcal{S}} (W+g) \qquad \textbf{(Init. of the algorithm)}$$

$$+ v_{\max} \sum_{t=1}^{k^{\mathrm{out}}} \gamma_{t,0}\, k_t^{\mathrm{in}}\, \mathbb{E}\left[\|\mathcal{E}_t\|^2\right] + v_{\max} \sum_{t=1}^{k^{\mathrm{out}}} \sum_{k=1}^{k_t^{\mathrm{in}}} \left(k_t^{\mathrm{in}} - k + 1\right) \gamma_{t,k}\, \mathcal{U}_{t,k} \;,$$

(Init. of the control variates) (Approximation of the h_i's)

where

$$\mathcal{E}_t := \mathsf{S}_{t,0} - \bar{\mathsf{h}}(\hat{S}_{t,0}) \qquad \mathcal{U}_{t,k} := \frac{2\, C_b}{m_{t,k}} + \frac{C_b^2}{m_{t,k}^2} + \frac{C_v}{\mathsf{b}\, M_{t,k}} + \frac{2\, C_{vb}}{\sqrt{\mathsf{b}}\, \bar{M}_{t,k}} + \frac{C_{vb}^2}{\mathsf{b}\, \bar{M}_{t,k}^2} \;.$$

Key ingredient for the proof: Lyapunov function

- The classical proof does not work

$$W(s_{t+1}) \leq W(s_t) + \langle \nabla W(s_t), s_{t+1} - s_t \rangle + \frac{L_{\dot{W}}}{2} \|s_{t+1} - s_t\|^2$$

Key ingredient for the proof: Lyapunov function

- The classical proof does not work

$$\mathbb{E}\left[W(s_{t+1})\middle|\mathcal{F}_t\right] \leq W(s_t) + \left\langle \nabla W(s_t), \mathbb{E}\left[s_{t+1} - s_t \middle| \mathcal{F}_t\right]\right\rangle + \frac{L_{\dot{W}}}{2}\mathbb{E}\left[\|s_{t+1} - s_t\|^2\middle|\mathcal{F}_t\right]$$

$$\mathbb{E}\left[s_{t+1} - s_t\middle|\mathcal{F}_t\right] = -\gamma_{t+1}\,\nabla W(s_t)$$ not true in our case

Key ingredient for the proof: Lyapunov function

- The classical proof does not work

$$\mathbb{E}\left[W(s_{t+1})\middle|\mathcal{F}_t\right] \leq W(s_t) + \left\langle \nabla W(s_t), \mathbb{E}\left[s_{t+1} - s_t\middle|\mathcal{F}_t\right]\right\rangle + \frac{L_{\dot{W}}}{2}\mathbb{E}\left[\|s_{t+1} - s_t\|^2\middle|\mathcal{F}_t\right]$$

$$\mathbb{E}\left[s_{t+1} - s_t\middle|\mathcal{F}_t\right] = -\gamma_{t+1}\,\nabla W(s_t)$$

not true in our case

- In our case:

$$s_{t+1} - s_t = \mathrm{prox}_{\gamma_{t+1}g}^{\mathsf{B}(s_t)}\left(s_t + \gamma_{t+1}\mathsf{S}_{t+1}\right) - s_t$$

$$\mathbb{E}\left[\mathsf{S}_{t+1}\middle|\mathcal{F}_t\right] \neq \mathsf{h}(s_t) \qquad \mathsf{h}(s_t) := -\mathsf{B}^{-1}(s_t)\,\frac{1}{n}\sum_{i=1}^{n} G_i(s_t)$$

Key ingredient for the proof: Lyapunov function

- The classical proof does not work

$$\mathbb{E}\left[W(s_{t+1})\middle|\mathcal{F}_t\right] \leq W(s_t) + \left\langle \nabla W(s_t), \mathbb{E}\left[s_{t+1} - s_t\middle|\mathcal{F}_t\right]\right\rangle + \frac{L_{\dot{W}}}{2}\mathbb{E}\left[\|s_{t+1} - s_t\|^2\middle|\mathcal{F}_t\right]$$

$$\mathbb{E}\left[s_{t+1} - s_t\middle|\mathcal{F}_t\right] = -\gamma_{t+1}\nabla W(s_t)$$ not true in our case

- In our case:

$$s_{t+1} - s_t = \mathrm{prox}_{\gamma_{t+1}\,g}^{\mathsf{B}(s_t)}\left(s_t + \gamma_{t+1}\mathsf{S}_{t+1}\right) - s_t$$

$$\mathbb{E}\left[\mathsf{S}_{t+1}\middle|\mathcal{F}_t\right] \neq \mathsf{h}(s_t) \qquad \mathsf{h}(s_t) := -\mathsf{B}^{-1}(s_t)\,\frac{1}{n}\sum_{i=1}^{n} G_i(s_t)$$

- Another strategy for the Lyapunov function F. and Moulines (2022, Lemma 7.9. and Proposition 7.10)

$$\begin{aligned}\mathbb{E}\left[W(s_{t+1}) + g(s_{t+1})\middle|\mathcal{F}_t\right] &\leq W(s_t) + g(s_t) \\ &\quad - \gamma_{t+1}(1/2 + o(\gamma_{t+1}))\,\mathbb{E}\left[\Delta^\star_{t+1} + \mathcal{D}^\star_{t+1}\right] \\ &\quad + \gamma_{t+1}\mathbb{E}\left[\|\mathsf{S}_{t+1} - \mathsf{h}(s_t)\|^2_{\mathsf{B}(s_t)}\right]\end{aligned}$$

Coro 1. The stepsize sequence

- Sufficient conditions :

- Constant when exact h_i's or randomly approximated with no bias
- Decreasing when deterministic approximation or randomly approximated with bias

$h_i(s)$'s	No approx Random, Unbiased i.e. $C_b = 0$	Determ. approx Random, Biased i.e. $C_b > 0$
	$\gamma_\star$	$\gamma_{t,k} \downarrow$, $\gamma_\star > \max \gamma_{t,k}$

$$\gamma_{t,k+1} := \gamma_{t,0} \prod_{j=0}^{k} \left(1 + \frac{2C_b}{m_{t,j+1}}\right)^{-1}$$

where

$$\gamma_{t,0} < \frac{1}{4L\upsilon_{\max}\upsilon} \frac{\mathsf{b}}{k_t^{\mathrm{in}}} \left(\sqrt{\frac{L_{\dot{W}}^2}{L^2} + 4\upsilon_{\min}\upsilon_{\max}\frac{k_t^{\mathrm{in}}}{\mathsf{b}}\upsilon} - \frac{L_{\dot{W}}}{L}\right).$$

Coro 2. Exact h_i's (i.e. $\mathcal{U}_t = 0$) and $\mathcal{E}_t = 0$ and $k_t^{\mathrm{in}} = k^{\mathrm{in}}$

In order to satisfy

$$\mathbb{E}\left[\Delta_\tau^\star\right] \leq \epsilon \qquad \tau \sim \mathcal{U}\left(\{1, \cdots, k^{\mathrm{out}}\} \times \{1, \cdots, k^{\mathrm{in}}\}\right)$$

- Stepsize sequence : $\gamma_\star = \frac{v_{\min}}{4L_{\dot{W}}}$ independent of ϵ
- Size of the minibatches, nbr of inner loops, nbr of outer loops

$$\mathsf{b} = O\left(\sqrt{n}\, v_{\min} v_{\max} \frac{L}{L_{\dot{W}}}\right) \qquad k^{\mathrm{in}} = O\left(\sqrt{n} \frac{L_{\dot{W}}}{L}\right) \qquad k^{\mathrm{out}} = O\left(\frac{1}{\epsilon \sqrt{n}} \frac{L}{v_{\min}}\right)$$

- Nbr of proximal steps and Nbr of calls to h_i

$$\mathcal{K}_{\mathrm{prox}} = O\left(\frac{1}{\epsilon} \frac{L_{\dot{W}}}{v_{\min}}\right) \qquad \mathcal{K}_{\bar{\mathsf{h}}} = O\left(\frac{\sqrt{n}}{\epsilon} L \frac{\sqrt{v_{\max}}}{\sqrt{v_{\min}}}\right)$$

In adequation with the literature when 3P-SPIDER $\equiv$ Precond Proximal-Gdt Wang et al (2019)

Complete the literature when 3P-SPIDER $\equiv$ incremental EM Fort et al (2020)

Coro 3. Unbiased Monte Carlo approximation of the h_i's

What is the cost of inexact preconditioned forward operators ?

- By choosing

$$\mathbb{E}\left[\|\mathcal{E}_t\|^2\right] = O\left(\frac{\epsilon^{1-\mathsf{a}'}}{(\sqrt{n}t)^{\mathsf{a}'}}\right), \qquad M_{t,k+1} = O\left(\frac{n^{(\mathsf{a}-\bar{\mathsf{a}})/2}}{\epsilon^{1-\mathsf{a}}}\, t^{\mathsf{a}}\,(k+1)^{\bar{\mathsf{a}}}\right)$$

for some $\mathsf{a}', \mathsf{a}, \bar{\mathsf{a}} \in [0,1)$

- then,

the same rates as with exact h_i's, at the price of a Monte Carlo complexity

$$\mathcal{K}_{\mathrm{MC}} = O\left(\frac{\sqrt{n}}{\epsilon^2}\right) \qquad \text{whatever } \mathsf{a}', \mathsf{a}, \bar{\mathsf{a}} \in [0,1)$$

VI. Bibliography

Bibliography (1/3)

Allen-Zhu, Z. and Hazan, E. *Variance reduction for faster non-convex optimization*, IMLS, 2016.

Allen-Zhu, Z. *Natasha 2: Faster Non-Convex Optimization Than SGD*, NeurIPS, 2018.

Atchadé, Y.F. and Fort, G. and Moulines, E. *On Perturbed Proximal Gradient Algorithms*, Journal of Machine Learning Research, 2017.

Cappé, O. and Moulines, E. *On-line Expectation Maximization algorithm for latent data models*, J. Roy. Stat. Soc. B Met., 2009.

Celeux, G. and Diebolt, J. *The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem*, Computational Statistics Quarterly 1985.

Chen, H.H.-G. and Rockafellar, R.T. *Convergence Rates in Forward-Backward Splitting*, SIAM J. Optim., 1997.

Chen, J. and Zhu, J. and Teh, Y.W. and Zhang, T. *Stochastic Expectation Maximization with Variance Reduction*, NeurIPS, 2018.

Chouzenoux, E. and Pesquet, J.-C. and Repetti, A. *Variable Metric Forward–Backward Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function*, Journal of Optimization Theory and Applications, 2014.

Combettes, P.L. and Vũ, B.C. *Variable metric forward–backward splitting with applications to monotone inclusions in duality*, Optimization, 2014.

Delyon, B. and Lavielle, M. and Moulines, E. *Convergence of a Stochastic Approximation version of the EM algorithm*, Ann. Statist., 1999.

Dempster, A.P. and Laird, N.M. and Rubin, D.B. *Maximum Likelihood from Incomplete Data via the EM Algorithm*, J. Roy. Stat. Soc. B Met., 1977.

Dongruo, Z. and Pan, X. and Quanquan, G. *Stochastic Nested Variance Reduction for Nonconvex Optimization*, Journal of Machine Learning Research, 2020.

Fang, C. and Li, C.J. and Lin, Z. and Zhang, T. *SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator*, NeurIPS, 2018.

Fort, G. and Gach, P. and Moulines, E. *Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence*, Stat. Comput., 2021.

Fort, G. and Moulines, E. *Convergence of the Monte Carlo Expectation Maximization for curved exponential families*, Ann. Statist., 2003.

Bibliography (2/3)

Fort, G. and Moulines, E. and Wai, H.-T. *A Stochastic Path-Integrated Differential Estimator Expectation Maximization Algorithm*, NeurIPS, 2020.

Ghadimi, S. and Lan, G. *Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming*, SIAM Journal on Optimization, 2013.

Ghadimi, S. and Lan, G. and Zhang, H. *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization*, Math. Program., 2016.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex Analysis and Minimization Algorithms*, 1996.

Karimi, B. and Wai, H.-T. and Moulines, E. and Lavielle, M. *On the Global Convergence of (Fast) Incremental Expectation Maximization Methods*, NeurIPS, 2019.

Karimi, H. and Nutini, J. and Schmidt, M. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, Machine Learning and Knowledge Discovery in Databases, 2016.

Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*, Springer International Publishing, 2020.

Li, Z. and Li, J. *A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization*, NeurIPS, 2018.

Moreau, J.J. *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société Mathématique de France, 1965.

McLachlan, G. J. and Krishnan, T. *The EM algorithm and extensions*, 2008.

Metel, M.R. and Takeda, A. *Stochastic Proximal Methods for Non-Smooth Non-Convex Constrained Sparse Optimization*, Journal of Machine Learning Research, 2021.

Neal, R. M. and Hinton, G. E. *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants* Learning in Graphical Models, 1998.

Nhan, H. P. and Lam, M. N. and Dzung, T. P. and Quoc, T.-D. *ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization*, Journal of Machine Learning Research, 2020.

Ng, S. K. and McLachlan, G. J. *On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures*, Stat. Comput., 2003.

Bibliography (3/3)

Nguyen, L.N. and Liu, J. and Scheinberg, K. and Takáč, M. *SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient*, ICML, 2017.

Reddi, S. J. and Hefny, A. and Sra, S. and Poczos, B. and Smola, A. *Stochastic Variance Reduction for Nonconvex Optimization*, ICML 2016.

Repetti, A. and Chouzenoux, E. and Pesquet, J.-C. *A preconditioned Forward-Backward approach with application to large-scale nonconvex spectral unmixing problems*, ICASSP 2014.

Wang, Z. and Ji, K. and Zhou, Y. and Liang, Y. and Tarokh, V. *SpiderBoost and Momentum: Faster Variance Reduction Algorithms*, NeurIPS, 2019.

Wei, G.C.G. and Tanner, M.A. *A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms*, J. Am. Stat. Assoc., 1990.

Wu, C.F.J. *On the Convergence Properties of the EM Algorithm*, Ann. Statist., 1983.

Zhang, J. and Xiao, L. *A Stochastic Composite Gradient Method with Incremental Variance Reduction*, NeurIPS, 2019.