Nonconvex Optimization for Latent Data Models: An Incremental and An Online Point of View

Samsung-HSE Laboratory, National Research University Higher School of Economics

Belhal Karimi

CMAP (Ecole Polytechnique) & XPOP (INRIA Saclay)

July 24, 2019

Overview

1. Statistical Learning in Latent Data Models

Supervised Learning

- Given input-output pair of random variables (X, Y) taking values in arbitrary input set $\mathsf{X} \subset \mathbb{R}^p$ and arbitrary output set $\mathsf{Y} \subset \mathbb{R}^q$ from unknown distribution $\mathcal{P}$.
- *Modeling* phase: $M_{\boldsymbol{\theta}} : \mathsf{X} \mapsto \mathsf{Y}$ of parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, called the *predictor*.
- Performance measured using a *loss* function $\ell : \mathsf{Y} \times \mathsf{Y} \mapsto \mathbb{R}$ where $\ell(y, y')$ is the loss incurred when the true output is y whereas y' is predicted.
- *Training* phase boils down to computing the following quantity:

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \overline{\mathcal{L}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \mathcal{L}(\boldsymbol{\theta}) + \mathsf{R}(\boldsymbol{\theta}) \right\} \tag{1}$$

with

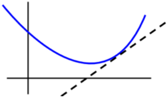
$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y) \sim \mathcal{P}} \left[\ell(y, M_{\boldsymbol{\theta}}(x)) \right] \quad \text{or} \quad \mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell(y_i, M_{\boldsymbol{\theta}}(x_i)) \,. \tag{2}$$

Nonconvex Optimization

$\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$

Convex function

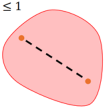
$f(\lambda w_1 + (1-\lambda) w_2) \le \lambda f(w_1) + (1-\lambda) f(w_2)$,
$0 \le \lambda \le 1$

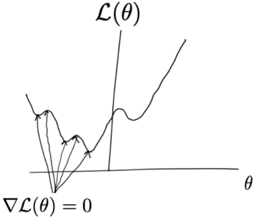


$\mathcal{C} \subseteq \mathbb{R}^d$

Convex set

$\forall w_1, w_2 \in C, \lambda w_1 + (1-\lambda) w_2 \in C$
$0 \le \lambda \le 1$





- For convex problems, we generally use $|\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)|$ (or $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$) as stability condition. where $\boldsymbol{\theta}^*$ is the optimal solution that can efficiently be found in the convex case.
- Yet, in the nonconvex case we use $\|\nabla \mathcal{L}(\boldsymbol{\theta})\|^2$, as advocated in (Nesterov, 2004) and (Ghadimi and Lan, 2013).
- A point $\boldsymbol{\theta}^*$ is said to be ε-stationary if $\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|^2 \le \varepsilon$. A stochastic iterative algorithm is said to achieve ε-stationarity in $\Gamma > 0$ iterations if $\mathbb{E}\left[\left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{(R)}) \right\|^2 \right] \le \varepsilon$.

Latent Data Models

- Models where the input-output relationship is not completely characterized by the observed $(x, y) \in \mathsf{X} \times \mathsf{Y}$ pairs in the training set
- Dependence on a set of unobserved latent variables $z \in \mathsf{Z} \subset \mathbb{R}^m$.
- Mandatory: Simulation step to complete the observed data with realizations of the latent variables.
- Formally, this specificity in our setting implies extending the loss function ℓ to accept a third argument as follows:

$$\ell(y, M_{\boldsymbol{\theta}}(x)) = \int_{\mathsf{Z}} \ell(z, y, M_{\boldsymbol{\theta}}(x)) dz \; . \tag{3}$$

Some Examples of Latent Data Models

- Include the incomplete data framework, *i.e.*, some observations are missing, but are far broader than that: for example, the latent structure could stem from the unknown labels in mixture models or hidden states in Hidden Markov Models.
- **Missing Data**: y stands for the observed data and the latent variables z are the missing data.
- **Mixed Effects Models**: the latent variables z are the random effects and identifying the structure of the latent data mainly corresponds to the inter-individual variability among the individuals of the dataset.
- **Mixture Models**: the latent variables correspond to the unknown mixture labels taking values in a discret finite set.

Quick Overview

- *L-Smoothness*: A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz-continuous, *i.e.*, for all $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$:

$$\|\nabla f(\theta) - \nabla f(\vartheta)\| \leq L \|\theta - \vartheta\| \ . \tag{4}$$

- We will deal with constrained $(\arg\min_{\boldsymbol{\theta}\in\Theta} \overline{\mathcal{L}}(\boldsymbol{\theta}))$ and unconstrained problems.

Algorithm	Gradient	Non-gradient	MC	Step.
SGD	$\mathcal{O}(1/\varepsilon^2)$ (Ghadimi and Lan, 2013)	?	x	γ_k
GD	$\mathcal{O}(n/\varepsilon)$ (Nesterov, 2004)	?	x	γ
SVRG/SAGA	$\mathcal{O}(n^{2/3}/\varepsilon)$ (Reddi et al., 2016)	$\mathcal{O}(n^{2/3}/\varepsilon)$ (Karimi et al., 2019c)	x	γ_k
MISO	$\mathcal{O}(n/\varepsilon)$ (Karimi et al., 2019b)	$\mathcal{O}(n/\varepsilon)$ (Karimi et al., 2019b)	x	—
MISSO	$\mathcal{O}(n/\varepsilon)$ (Karimi et al., 2019b)	$\mathcal{O}(n/\varepsilon)$ (Karimi et al., 2019b)	✓	—
Biased SA	$\mathcal{O}(c_0 + \frac{\log(n)}{\varepsilon\sqrt{n}})$ (Karimi et al., 2019a)	$\mathcal{O}(c_0 + \frac{\log(n)}{\varepsilon\sqrt{n}})$ (Karimi et al., 2019a)	✓	γ_k

Table 1: ERM methods: Table comparing the complexity, measured in terms of iterations, of different algorithms for non-convex optimization. MC stands for Monte Carlo integration of the drift term and Step. for stepsize.

2. Nonconvex Risk Minimization

2. Nonconvex Risk Minimization

2.1 Incremental Method for Non-smooth Nonconvex Objective: with Applications to Logistic Regression and Bayesian Deep Learning

Large-scale machine learning

Constrained Minimization of large sum of functions

We are interested in the minimization of a large finite-sum of functions:

$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) , \tag{5}$$

where Θ is a convex, compact, and closed subset of $\mathbb{R}^p$, and for any $i \in [\![1, n]\!]$, the function $\mathcal{L}_i : \mathbb{R}^p \to \mathbb{R}$ is bounded from below and is (possibly) nonconvex and non-smooth.

Some examples Given data points $(x_i, i \in [\![1, n]\!])$ and observations $(y_i, i \in [\![1, n]\!])$

- Maximum likelihood estimation: $\mathcal{L}(\boldsymbol{\theta}) \triangleq -\sum_{i=1}^{N} \log p_i(y_i, \boldsymbol{\theta})$
- Variational inference: $\mathcal{L}(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \mathrm{KL}\left(q_i(w; \boldsymbol{\theta}) || \, p_i(w | y_i, x_i)\right)$
- Logistic regression: $\mathcal{L}(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \log(1 + e^{-y_i < \boldsymbol{\theta}, x_i >})$

Majorization-Minimization principle

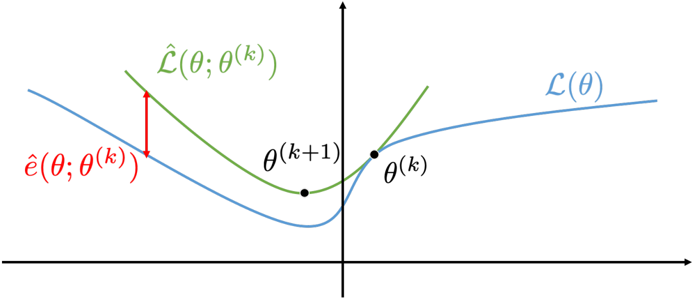


Figure 1: Majorization-Minimization principle

- Iteratively minimize locally tight upper bounds on the objective
- Drives the objective function downwards
- Examples: the proximal gradient algorithm (Beck and Teboulle, 2009), the EM algorithm (McLachlan and Krishnan, 2007) and variational inference (Wainwright and Jordan, 2008).

Notations and Assumptions

- Constrained optimization: Θ convex subset of $\mathbb{R}^p$.
- For all $i \in [\![1, n]\!]$, $\mathcal{L}_i$ is continuously differentiable on Θ.
- For all $i \in [\![1, n]\!]$, $\mathcal{L}_i$ is bounded from below, i.e. there exist a constant $M_i \in \mathbb{R}$ such as for all $\boldsymbol{\theta} \in \Theta$, $\mathcal{L}_i(\boldsymbol{\theta}) \geq M_i$.
- $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \cdot) : \mathbb{R}^p \to \mathbb{R}$ is a surrogate of $\mathcal{L}_i$ at $\boldsymbol{\theta}$ if the following properties are satisfied:
  1. the function $\vartheta \to \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \vartheta)$ is continuously differentiable on Θ
  2. for all $\vartheta \in \Theta$, $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \vartheta) \geq \mathcal{L}_i(\vartheta)$, $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}) = \mathcal{L}_i(\boldsymbol{\theta})$ and $\nabla \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \vartheta)\Big|_{\vartheta = \boldsymbol{\theta}} = \nabla \mathcal{L}_i(\vartheta)\Big|_{\vartheta = \boldsymbol{\theta}}$.

Incremental Surrogate Minimization

The incremental scheme of (Mairal, 2015) computes surrogate functions, at each iteration of the algorithm, for a mini-batch of components:

Algorithm 1 MISO algorithm

Initialization: given an initial parameter estimate $\hat{\boldsymbol{\theta}}^{(0)}$, for all $i \in [\![1, n]\!]$ compute a surrogate function $\vartheta \rightarrow \widehat{\mathcal{L}}_i(\hat{\boldsymbol{\theta}}^{(0)}; \vartheta)$.

Iteration k: given the current estimate $\hat{\boldsymbol{\theta}}^{(k)}$:

1. Pick i_k uniformly from $[\![1, n]\!]$.
2. Update $\mathcal{A}_i^{k+1}(\boldsymbol{\theta})$ as:

$$\mathcal{A}_i^{k+1}(\boldsymbol{\theta}) = \begin{cases} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}), & \text{if } i = i_k \\ \mathcal{A}_i^k(\boldsymbol{\theta}), & \text{otherwise.} \end{cases}$$

3. Set $\hat{\boldsymbol{\theta}}^{(k+1)} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{A}_i^{k+1}(\boldsymbol{\theta})$.

Intractable surrogate functions

- In many cases of interest those surrogates are intractable.
- Denote by $z = (z_i \in \mathsf{Z}, i \in [\![1, n]\!])$ where Z is a subset of $\mathbb{R}^{m_i}$ as set of latent variables.
- For all $i \in [\![1, n]\!]$, let μ_i be a σ-finite measure on the Borel σ-algebra $\mathcal{Z} = \mathcal{B}(\mathsf{Z})$.
- $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ be a family of probability densities with respect to μ_i, and $r_{i,\theta} : \mathsf{Z} \times \Theta \to \mathbb{R}$ be functions such that:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \int_{\mathsf{Z}} r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i) p_i(z_i; \overline{\boldsymbol{\theta}}) \mu_i(dz_i) \quad \forall \, (\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) \in \Theta \times \Theta \, . \quad (6)$$

The surrogate function denoted $\widehat{\mathcal{L}}_i(\theta; \vartheta)$ is fully defined by the pair $(r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i), p_i(z_i, \overline{\boldsymbol{\theta}}))$.

Examples of intractable surrogates

Incremental EM algorithm

- In the missing data context, let $c_i(z_i, \boldsymbol{\theta})$ be the joint likelihood of the observations and the latent data referred to as the complete likelihood.
- $g_i(\boldsymbol{\theta}) \triangleq \int_{\mathsf{Z}} c_i(z_i, \boldsymbol{\theta}) \mu_i(z_i)$ is the likelihood of the observations (in which the latent variables are marginalized).

The incremental EM algorithm falls into the incremental MM framework:

- For $i \in [\![1, n]\!]$ and $\boldsymbol{\theta} \in \Theta$ the loss function $\ell_i(\boldsymbol{\theta}) \triangleq -\log g_i(\boldsymbol{\theta})$
- for $\vartheta \in \Theta$ the surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \vartheta)$, introduced in the pioneering paper (Neal and Hinton, 1998), is defined by

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \vartheta) \triangleq \int_{\mathsf{Z}} \log \frac{p_i(z_i, \boldsymbol{\theta})}{c_i(z_i, \vartheta)} p_i(z_i, \boldsymbol{\theta}) \mu_i(z_i) = \mathrm{KL}\left(p_i(z_i, \boldsymbol{\theta}) \,||\, p_i(z_i, \vartheta)\right) + \ell_i(\vartheta) \tag{7}$$

In most cases, this surrogate is intractable.

Examples of intractable surrogates

Variational Inference Let $x = (x_i, i \in [\![1, n]\!])$ and $y = (y_i, i \in [\![1, n]\!])$ be i.i.d. input-output pairs and w be a global latent variable taking values in W a subset of $\mathbb{R}^J$.

A natural decomposition of the joint distribution is:

$$p(y, w|x) = \pi(w) \prod_{i=1}^n p(y_i|x_i, w) \ . \tag{8}$$

The variational inference problem boils down to minimizing the following KL divergence:

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \Theta} \ \mathcal{L}(\boldsymbol{\theta}) &:= \mathrm{KL}\left(q(w; \boldsymbol{\theta}) \,||\, p(w|y, x)\right) \\ &:= \mathbb{E}_{q(w;\boldsymbol{\theta})}\big[\log\left(q(w;\boldsymbol{\theta})/p(w|y,x)\right)\big] \ . \end{aligned} \tag{9}$$

Using (8), we decompose $\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}) + \mathrm{const.}$ where:

$$\begin{aligned} \mathcal{L}_i(\boldsymbol{\theta}) := &- \mathbb{E}_{q(w;\boldsymbol{\theta})}\big[\log p(y_i|x_i, w)\big] \\ &+ \frac{1}{n}\mathbb{E}_{q(w;\boldsymbol{\theta})}\big[\log q(w;\boldsymbol{\theta})/\pi(w)\big] = r_i(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta}) \ . \end{aligned} \tag{10}$$

Examples of intractable surrogates

Variational Inference

- MISSO method with a quadratic surrogate function defined as:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}) := \mathcal{L}_i(\overline{\boldsymbol{\theta}}) + \left\langle \nabla_{\boldsymbol{\theta}}\mathcal{L}_i(\overline{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{\mathrm{L}}{2}\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \,. \qquad (11)$$

- Let $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$ be a differentiable function *w.r.t.* $\boldsymbol{\theta} \in \Theta$ s.t. $w = t(z, \overline{\boldsymbol{\theta}})$, where $z \sim \mathcal{N}_d(0, I)$, is distributed according to $q(\cdot, \overline{\boldsymbol{\theta}})$.
- By (Blundell et al., 2015, Proposition 1), the gradient of $-r_i(\cdot)$ in (10) is (re-parametrization trick):

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q(w;\overline{\boldsymbol{\theta}})}\left[\log p(y_i|x_i, w)\right] = \mathbb{E}_{z\sim\mathcal{N}_d(0,I)}\left[\mathrm{J}^t_{\boldsymbol{\theta}}(z,\overline{\boldsymbol{\theta}})\nabla_w \log p(y_i|x_i,w)\big|_{w=t(z,\overline{\boldsymbol{\theta}})}\right] .$$

- For most cases, the term $\nabla\,\mathrm{R}(\overline{\boldsymbol{\theta}})$ can be evaluated in closed form.

$$r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z) := \left\langle \nabla_{\boldsymbol{\theta}} d(\overline{\boldsymbol{\theta}}) - \mathrm{J}^t_{\boldsymbol{\theta}}(z,\overline{\boldsymbol{\theta}})\nabla_w \log p(y_i|x_i,w)\big|_{w=t(z,\overline{\boldsymbol{\theta}})} \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|^2 \,.$$

Proposed Method: MISSO

- *Minimization by Incremental Stochastic Surrogate Optimization* (MISSO) method: expectation approximated by *Monte Carlo* integration.
- Denote by $M \in \mathbb{N}$ the Monte Carlo batch size and let $z_m \in \mathsf{Z}$, $m = 1, ..., M$ be a set of samples. These samples can be drawn (Case 1) i.i.d. from the distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$ or (Case 2) from a Markov chain with the stationary distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$.
- We define the following stochastic surrogate:

$$\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M) := \frac{1}{M} \sum_{m=1}^{M} r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_m) \tag{12}$$

MISSO Algorithm

Algorithm 2 MISSO algorithm

Initialization: $\hat{\boldsymbol{\theta}}^{(0)}$; a sequence of non-negative numbers $\{M_{(k)}\}_{k=0}^{\infty}$.

For all $i \in [\![1, n]\!]$, draw $M_{(0)}$ samples from $p_i(\cdot; \hat{\boldsymbol{\theta}}^{(0)})$ and $\widetilde{\mathcal{A}}_i^0(\boldsymbol{\theta}) := \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(k)}})$.

Iteration k: given the current estimate $\hat{\boldsymbol{\theta}}^{(k)}$:

1. Pick a function index i_k uniformly on $[\![1, n]\!]$.
2. Draw $M_{(k)}$ Monte-Carlo samples from $p_i(\cdot; \hat{\boldsymbol{\theta}}^{(k)})$.
3. Update the individual surrogate functions recursively as:

$$\widetilde{\mathcal{A}}_i^{k+1}(\boldsymbol{\theta}) = \begin{cases} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}, \{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}), & \text{if } i = i_k \\ \widetilde{\mathcal{A}}_i^{k}(\boldsymbol{\theta}), & \text{otherwise.} \end{cases} \tag{13}$$

4. Set $\hat{\boldsymbol{\theta}}^{(k+1)} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^{k+1}(\boldsymbol{\theta})$.

Convergence analysis: Assumptions

(S1) For all $i \in [\![1, n]\!]$ and $\overline{\boldsymbol{\theta}} \in \Theta$, the function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ is convex *w.r.t.* $\boldsymbol{\theta}$, and it holds

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta}), \ \forall \ \boldsymbol{\theta} \in \Theta \ , \tag{14}$$

where the equality holds when $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$.

(S2) For any $\overline{\boldsymbol{\theta}}_i \in \Theta$, $i \in [\![1, n]\!]$ and some $\epsilon > 0$, the difference function $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}_i) - \mathcal{L}(\boldsymbol{\theta})$ is defined for all $\boldsymbol{\theta} \in \Theta_\epsilon$ and differentiable for all $\boldsymbol{\theta} \in \Theta$, where $\Theta_\epsilon = \{\boldsymbol{\theta} \in \mathbb{R}^d, \inf_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \epsilon\}$ is an ϵ-neighborhood set of Θ. Moreover, for some constant L, the gradient satisfies

$$\|\nabla \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)\|^2 \leq 2L \, \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n), \ \forall \ \boldsymbol{\theta} \in \Theta \ . \tag{15}$$

Convergence analysis: Assumptions

(H1) For all $i \in [\![1, n]\!]$, $\overline{\boldsymbol{\theta}} \in \Theta$, $z_i \in \mathsf{Z}$, the measurable function $r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i)$ is convex in $\boldsymbol{\theta}$ and is lower bounded.

(H2) For the samples $\{z_{i,m}\}_{m=1}^{M}$, there exists finite constants C_{r} and C_{gr} such that

$$C_{\mathrm{r}} := \sup_{\overline{\boldsymbol{\theta}} \in \Theta} \sup_{M>0} \frac{1}{\sqrt{M}} \mathbb{E}_{\overline{\boldsymbol{\theta}}} \left[\sup_{\boldsymbol{\theta} \in \Theta} \left| \sum_{m=1}^{M} \left\{ r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_{i,m}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \right\} \right| \right] \tag{16}$$

$$C_{\mathrm{gr}} := \sup_{\overline{\boldsymbol{\theta}} \in \Theta} \sup_{M>0} \sqrt{M} \mathbb{E}_{\overline{\boldsymbol{\theta}}} \left[\sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{M} \sum_{m=1}^{M} \frac{\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}) - r_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}, z_{i,m})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \right|^2 \right] \tag{17}$$

for all $i \in [\![1, n]\!]$, and we denoted by $\mathbb{E}_{\overline{\boldsymbol{\theta}}}[\cdot]$ the expectation *w.r.t.* a Markov chain $\{z_{i,m}\}_{m=1}^{M}$ with initial distribution $\xi_i(\cdot; \overline{\boldsymbol{\theta}})$, transition kernel $P_{i,\overline{\boldsymbol{\theta}}}$, and stationary distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$.

Constrained Optimization

- As problem (5) is a constrained optimization, we consider the following stationarity measure:

$$g(\overline{\boldsymbol{\theta}}) := \inf_{\boldsymbol{\theta} \in \Theta} \frac{\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \quad \text{and} \quad g(\overline{\boldsymbol{\theta}}) = g_+(\overline{\boldsymbol{\theta}}) - g_-(\overline{\boldsymbol{\theta}}) \,, \quad (18)$$

  where $g_+(\overline{\boldsymbol{\theta}}) := \max\{0, g(\overline{\boldsymbol{\theta}})\}$, $g_-(\overline{\boldsymbol{\theta}}) := -\min\{0, g(\overline{\boldsymbol{\theta}})\}$ denote the positive and negative part of $g(\overline{\boldsymbol{\theta}})$, respectively.
- Note that $\overline{\boldsymbol{\theta}}$ is a stationary point if and only if $g_-(\overline{\boldsymbol{\theta}}) = 0$ (Fletcher et al., 2002).
- Furthermore, suppose that the sequence $\{\hat{\boldsymbol{\theta}}^{(k)}\}_{k \geq 0}$ has a limit point $\overline{\boldsymbol{\theta}}$ that is a stationary point, then one has $\lim_{k \to \infty} g_-(\hat{\boldsymbol{\theta}}^{(k)}) = 0$.

Non-asymptotic analysis

Theorem 1

Under (S1), (S2), (H1), (H2). For any $K_{\max} \in \mathbb{N}$, *let* K *be an independent discrete r.v. drawn uniformly from* $\{0, ..., K_{\max} - 1\}$ *and define the following quantity:*

$$\Delta_{(K_{\max})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\hat{\boldsymbol{\theta}}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\max})}(\hat{\boldsymbol{\theta}}^{(K_{\max})})] + \sum_{k=0}^{K_{\max}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}}, \quad (19)$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}\left[\|\nabla \widehat{e}^{(K)}(\hat{\boldsymbol{\theta}}^{(K)})\|^2\right] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \quad (20)$$

$$\mathbb{E}[g_{-}(\hat{\boldsymbol{\theta}}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\mathsf{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}. \quad (21)$$

Non-asymptotic analysis

$$\Delta_{(K_{\max})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\hat{\boldsymbol{\theta}}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\max})}(\hat{\boldsymbol{\theta}}^{(K_{\max})})] + \sum_{k=0}^{K_{\max}-1} \frac{4LC_{\mathrm{r}}}{\sqrt{M_{(k)}}} , \quad (22)$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}\left[\|\nabla \widehat{e}^{(K)}(\hat{\boldsymbol{\theta}}^{(K)})\|^2\right] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \quad (23)$$

$$\mathbb{E}[g_{-}(\hat{\boldsymbol{\theta}}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\mathrm{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}. \quad (24)$$

- $\Delta_{(K_{\max})}$ is finite for any $K_{\max} \in \mathbb{N}$
- MISO method can be analyzed as a special case of the MISSO method satisfying $C_{\mathrm{r}} = C_{\mathrm{gr}} = 0$
- Then, Eq. (24) gives a non-asymptotic rate of $\mathbb{E}[g_{-}^{(K)}] \leq \mathcal{O}(\sqrt{nL/K_{\max}})$.

Asymptotic analysis

Under an additional assumption on the sequence of batch size $M_{(k)}$:

Theorem 2

Under (S1), (S2), (H1), (H2). In addition, assume that $\{M_{(k)}\}_{k\geq 0}$ *is a non-decreasing sequence of integers which satisfies* $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$. *Then:*

1. *the negative part of the stationarity measure converges almost surely to zero*, i.e., $\lim_{k\to\infty} g_-(\hat{\boldsymbol{\theta}}^{(k)}) = 0$ *a.s.*.
2. *the objective value* $\mathcal{L}(\hat{\boldsymbol{\theta}}^{(k)})$ *converges almost surely to a finite number* $\underline{\mathcal{L}}$, i.e., $\lim_{k\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}^{(k)}) = \underline{\mathcal{L}}$ *a.s.*.

In particular, the first result above shows that the sequence $\{\hat{\boldsymbol{\theta}}^{(k)}\}_{k\geq 0}$ produced by the MISSO method satisfies an *asymptotic stationary point condition*.

Logistic Regression with Missing Covariates

- $y = (y_i, i \in [\![1, n]\!])$ vector of binary responses and $z_i = (z_{i,p}) \in \mathbb{R}^p$ covariates.
- z_i is not fully observed. $z_{i,mis}$: missing values and $z_{i,obs}$: observed.
- $(z_i, i \in [\![1, n]\!][n]) \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Omega})$ where $\beta \in \mathbb{R}^p$ (i.i.d.)
- Model defined by

$$\operatorname{logit}(\mathbb{P}(y_{ij} = 0|z_i)) = d_{ij}^\top z_i$$

- Exponential family: Sufficient statistics are $\tilde{S}_i(z_i) \triangleq (z_i, z_i^\top z_i)$.
- MISSO algorithm consists in picking a set I_k, sampling a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ for $i \in I_k$ and computing the quantities $(s_i^{1,k}, s_i^{2,k})$ as follows:

$$(s_i^{1,k}, s_i^{2,k}) = \begin{cases} \left(\frac{1}{M_k}\sum_{m=0}^{M_k-1} z_i^{k,m}, \frac{1}{M_k}\sum_{m=0}^{M_k-1} (z_i^{k,m})^\top z_i^{k,m}\right) & \text{if } i \in I_k \\ (s_i^{1,k-1}, s_i^{2,k-1}) & \text{otherwise} \end{cases} \tag{25}$$

Then $\beta^k = \frac{1}{N}\sum_{i=1}^N s_i^{1,k}$ and $\Omega^k = \frac{1}{N}\sum_{i=1}^N s_i^{2,k} - (\beta^k)^\top \beta^k$

Logistic Regression on TraumaBase

- TraumaBase (http://traumabase.eu) dataset: 15 trauma centers in France, covering measurements on patients from the initial to last stage of trauma. (6384 and $p = 16$ quantitative variables of influence)
- Predict binary response: severe trauma or not.
- We apply the MISSO method to fit a logistic regression model

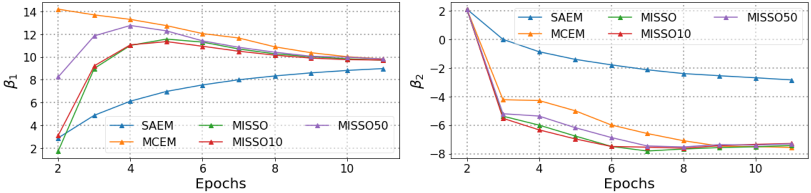


Figure 2: Convergence of two components of the vector of parameters β for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against the number of passes over the data.

Bayesian LeNet-5 on MNIST

layer type	width	stride	padding	input shape	nonlinearity
convolution (5×5)	6	1	0	$1 \times 32 \times 32$	ReLU
max-pooling (2×2)		2	0	$6 \times 28 \times 28$	
convolution (5×5)	6	1	0	$1 \times 14 \times 14$	ReLU
max-pooling (2×2)		2	0	$16 \times 10 \times 10$	
fully-connected	120			400	ReLU
fully-connected	84			120	ReLU
fully-connected	10			84	

Table 1: LeNet-5 architecture

- $N = 60\,000$ handwritten digits, 28×28 images, $d = 784$
- $p(w) = \mathcal{N}(0, I)$, $p(y_i|x_i, w) = \mathrm{Softmax}(f(x_i, w))$ where f is a NN.
- Variational distribution for layer j: $q(w_\ell, \theta_\ell)$ is a Gaussian distribution $\mathcal{N}(\mu_\ell, \sigma^2 I)$

Bayesian LeNet-5 on MNIST

- MISSO Updates:

$$\mu_\ell^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \sigma^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} , \tag{26}$$

where $\hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} = \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k-1)}$ and $\hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} = \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k-1)}$ for $i \neq i_k$ and:

$$\hat{\boldsymbol{\delta}}_{\mu_\ell,i_k}^{(k)} = -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} \nabla_w \log p(y_{i_k}|x_{i_k}, w)\Big|_{w=t(\hat{\boldsymbol{\theta}}^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} \mathrm{R}(\hat{\boldsymbol{\theta}}^{(k-1)}) ,$$

$$\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)} = -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} z_m^{(k)} \nabla_w \log p(y_{i_k}|x_{i_k}, w)\Big|_{w=t(\hat{\boldsymbol{\theta}}^{(k-1)}, z_m^{(k)})} + \nabla_{\sigma} \mathrm{R}(\hat{\boldsymbol{\theta}}^{(k-1)})$$

with $\mathrm{R}(\boldsymbol{\theta}) = n^{-1}\sum_{\ell=1}^{d}\left(-\log(\sigma) + (\sigma^2 + \mu_\ell^2)/2 - 1/2\right)$.

Bayesian LeNet-5 on MNIST

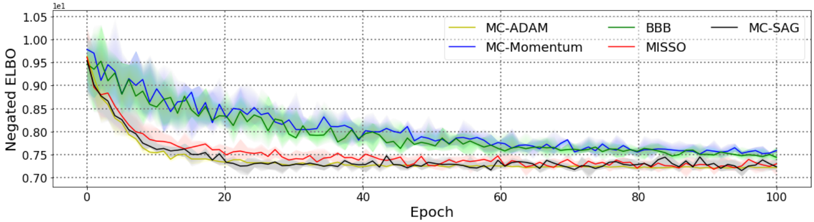


Figure 3: (Incremental Variational Inference) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

2. Nonconvex Risk Minimization

2.2 Online Optimization of Nonconvex Expected Risk: with Applications to Online and Reinforcement Learning

Stochastic Approximation (SA) Scheme

- Consider a smooth Lyapunov function $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ (possibly nonconvex) that we wish to find its *stationary point*.
- SA scheme (Robbins and Monro, 1951) is a stochastic process:

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n - \gamma_{n+1} H_{\boldsymbol{\eta}_n}(X_{n+1}), \quad n \in \mathbb{N}$$

  where $\boldsymbol{\eta}_n \in \mathcal{H} \subseteq \mathbb{R}^d$ is the nth state, $\gamma_n > 0$ is the step size.
- The *drift term* $H_{\boldsymbol{\eta}_n}(X_{n+1})$ depends on an **i.i.d. random element** X_{n+1} and the mean-field satisfies

$$h(\boldsymbol{\eta}_n) = \mathbb{E}\big[H_{\boldsymbol{\eta}_n}(X_{n+1})|\mathcal{F}_n\big] = \nabla V(\boldsymbol{\eta}_n),$$

  where $\mathcal{F}_n$ is the filtration generated by $\{\boldsymbol{\eta}_0, \{X_m\}_{m \le n}\}$.
- In this case, the SA scheme is better known as the SGD method.

Biased SA Scheme

In this work, we relax a few restrictions of the classical SA. Consider:

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n - \gamma_{n+1} H_{\boldsymbol{\eta}_n}(X_{n+1}), \quad n \in \mathbb{N}. \tag{27}$$

- The **mean field** $h(\boldsymbol{\eta}) \neq \nabla V(\boldsymbol{\eta})$

  $\Longrightarrow$ relevant to *non-gradient* method where the gradient is hard to compute, e.g., online EM.

- $\{X_n\}_{n\geq 1}$ is not i.i.d. and form a **state-dependent Markov chain**

  $\Longrightarrow$ relevant to *SGD with non-iid noise* and *policy gradient*. E.g., $\boldsymbol{\eta}_n$ controls the policy in a Markov decision process, and the gradient estimate $H_{\boldsymbol{\eta}_n}(x)$ is computed from the intermediate reward.

Biased SA Scheme

In this work, we relax a few restrictions of the classical SA. Consider:

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n - \gamma_{n+1} H_{\boldsymbol{\eta}_n}(X_{n+1}), \quad n \in \mathbb{N}. \tag{27}$$

- The **mean field** $h(\boldsymbol{\eta}) \neq \nabla V(\boldsymbol{\eta})$ but satisfies for some $c_0 \geq 0, c_1 > 0$,

$$c_0 + c_1 \left\langle \nabla V(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \right\rangle \geq \|h(\boldsymbol{\eta})\|^2$$

- $\{X_n\}_{n \geq 1}$ is not i.i.d. and form a **state-dependent Markov chain**:

$$\mathbb{E}[H_{\boldsymbol{\eta}_n}(X_{n+1}) | \mathcal{F}_n] = P_{\boldsymbol{\eta}_n} H_{\boldsymbol{\eta}_n}(X_n) = \int H_{\boldsymbol{\eta}_n}(x) P_{\boldsymbol{\eta}_n}(X_n, \mathrm{d}x),$$

  where $P_{\boldsymbol{\eta}_n} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$ is Markov kernel with a unique stationary distribution $\pi_{\boldsymbol{\eta}_n}$, and the mean field $h(\boldsymbol{\eta}) = \int H_{\boldsymbol{\eta}}(x) \pi_{\boldsymbol{\eta}}(\mathrm{d}x)$.

Prior Work & Biased SA Scheme

We consider two cases depending on the noise sequence

$$\boldsymbol{e}_{n+1} = H_{\boldsymbol{\eta}_n}(X_{n+1}) - h(\boldsymbol{\eta}_n)$$

Case 1: When $\{\boldsymbol{e}_n\}_{n\geq 1}$ is Martingale difference —

$$\mathbb{E}\big[\boldsymbol{e}_{n+1}|\mathcal{F}_n\big] = 0 \text{ and other conditions...}$$

- *Asymptotic Analysis*: with smooth $h(\cdot)$ (Robbins and Monro, 1951), (Benveniste et al., 1990), (Borkar, 2009).
- *Non-asymptotic Analysis*: focus on $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$,
  - Convex case: rate of $\mathcal{O}(1/n)$ in (Moulines and Bach, 2011), biased SA also studied in TD learning (Dalal et al., 2018).
  - Nonconvex case: (Ghadimi and Lan, 2013), (Bottou et al., 2018) studied convergence with martingale noise.

Prior Work & Biased SA Scheme

We consider two cases depending on the noise sequence

$$\boldsymbol{e}_{n+1} = H_{\boldsymbol{\eta}_n}(X_{n+1}) - h(\boldsymbol{\eta}_n)$$

Case 2: When $\{\boldsymbol{e}_n\}_{n \geq 1}$ is state-controlled Markov noise —

$$\mathbb{E}\big[\boldsymbol{e}_{n+1} | \mathcal{F}_n\big] = P_{\boldsymbol{\eta}_n} H_{\boldsymbol{\eta}_n}(X_n) - h(\boldsymbol{\eta}_n) \neq 0 \text{ and other conditions....}$$

- *Asymptotic Analysis*: studied with $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$ in (Kushner and Yin, 2003), similar biased SA setting in (Tadić and Doucet, 2017).
- *Non-asymptotic Analysis*: not many work here...
  - Sun et al. (2018) and Duchi et al. (2012) assumed $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$ & **state-independent** Markov chain.
  - Bhandari et al. (2018) studied a similar setting but focuses on linear SA with convex Lyapunov function.

Our Contributions

- First *non-asymptotic analysis* of biased SA scheme under the relaxed settings for *nonconvex* Lyapunov function.
- For both cases, with N being a r.v. drawn from $\{1, ..., n\}$, we show

$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] = \mathcal{O}\Big(c_0 + \frac{\log n}{\sqrt{n}}\Big)$$

  where c_0 is the *bias* of the mean field. If unbiased, then we find a stationary point.
- Analysis of two stochastic algorithms:
  - Online expectation maximization in (Cappé and Moulines, 2009)
  - Online policy gradient for infinite horizon reward maximization (Baxter and Bartlett, 2001).
- We provide the first *non-asymptotic* rates for the above algorithms.

General Assumptions

(A1) For all $\boldsymbol{\eta} \in \mathcal{H}$, there exists $c_0 \geq 0, c_1 > 0$ such that

$$c_0 + c_1 \langle \nabla V(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \rangle \geq \|h(\boldsymbol{\eta})\|^2$$

(A2) For all $\boldsymbol{\eta} \in \mathcal{H}$, there exists $d_0 \geq 0, d_1 > 0$ such that

$$d_0 + d_1 \|h(\boldsymbol{\eta})\| \geq \|\nabla V(\boldsymbol{\eta})\|$$

(A3) Lyapunov function V is L-smooth. For all $(\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{H}^2$,

$$\|\nabla V(\boldsymbol{\eta}) - \nabla V(\boldsymbol{\eta}')\| \leq L \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|$$

- **(A1)**, **(A2)** assume that the mean field $h(\boldsymbol{\eta})$ is indirectly related to the gradient of a Lyapunov function $V(\boldsymbol{\eta})$ ($h(\boldsymbol{\eta}) \neq \nabla V(\boldsymbol{\eta})$).
- If $c_0 = d_0 = 0$, then the SA scheme is *un-biased*.
- **(A3)** is the standard smoothness assumption.

Stopping Criterion

- We adopt a stopping rule similar to (Ghadimi and Lan, 2013) that is typical for *nonconvex* problems.
- Fix any $n \geq 1$ and let $N \in \{0, \ldots, n\}$ be a discrete random variable (independent of $\{\mathcal{F}_n,\ n \in \mathbb{N}\}$) with

$$\mathbb{P}(N = \ell) = \left(\sum_{k=0}^{n} \gamma_{k+1}\right)^{-1} \gamma_{\ell+1} \ , \tag{28}$$

  where N serves as the terminating iteration for (27).
- Throughout this talk, we assume N is distributed as (28) and study the estimator $\boldsymbol{\eta}_N$.

Case 1: Martingale Difference Noise

(A4) $\{\boldsymbol{e}_n\}_{n\geq 1}$ is a Martingale difference sequence such that $\mathbb{E}\left[\boldsymbol{e}_{n+1} \,|\, \mathcal{F}_n\right] = \mathbf{0}$, $\mathbb{E}\left[\|\boldsymbol{e}_{n+1}\|^2 \,\middle|\, \mathcal{F}_n\right] \leq \sigma_0^2 + \sigma_1^2 \|h(\boldsymbol{\eta}_n)\|^2$ for any $n \in \mathbb{N}$.

Theorem 3

Let A1, A3 and A4 hold and $\gamma_{n+1} \leq (2c_1 L(1+\sigma_1^2))^{-1}$ for all $n \geq 0$. Let $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$, we have

$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] \leq \frac{2c_1\left(V_{0,n} + \sigma_0^2 L \sum_{k=0}^n \gamma_{k+1}^2\right)}{\sum_{k=0}^n \gamma_{k+1}} + 2c_0\ ,$$

If we set $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$, then the SA scheme (27) finds an $\mathcal{O}(c_0 + \log n/\sqrt{n})$ quasi-stationary point within n iterations.

Case 1: Martingale Difference Noise

(A4) $\{\boldsymbol{e}_n\}_{n\geq 1}$ is a Martingale difference sequence such that $\mathbb{E}\left[\boldsymbol{e}_{n+1} \,|\, \mathcal{F}_n\right] = \mathbf{0}$, $\mathbb{E}\left[\|\boldsymbol{e}_{n+1}\|^2 \,\middle|\, \mathcal{F}_n\right] \leq \sigma_0^2 + \sigma_1^2 \|h(\boldsymbol{\eta}_n)\|^2$ for any $n \in \mathbb{N}$.

$\Longrightarrow$ can be satisfied when *X_n is i.i.d.* similar to the SGD setting.

Theorem 3

Let A1, A3 and A4 hold and $\gamma_{n+1} \leq (2c_1 L(1+\sigma_1^2))^{-1}$ for all $n \geq 0$. Let $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$, we have

$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] \leq \frac{2c_1\left(V_{0,n} + \sigma_0^2 L \sum_{k=0}^n \gamma_{k+1}^2\right)}{\sum_{k=0}^n \gamma_{k+1}} + 2c_0 \ ,$$

If we set $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$, then the SA scheme (27) finds an $\mathcal{O}(c_0 + \log n/\sqrt{n})$ quasi-stationary point within n iterations.

$\Longrightarrow$ if $h(\boldsymbol{\eta}) = \nabla V(\boldsymbol{\eta})$ it recovers *(Ghadimi and Lan, 2013, Theorem 2.1)*.

Case 2: State-dependent Markov Noise

In this case, $\{\boldsymbol{e}_n\}_{n\geq 1}$ is not a Martingale sequence. Instead,

(A5) There exists a Borel measurable function $\hat{H} : \mathcal{H} \times \mathsf{X} \to \mathcal{H}$,

$$\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) = H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}), \ \forall \ \boldsymbol{\eta} \in \mathcal{H}, x \in \mathsf{X}.$$

(A6) For all $\boldsymbol{\eta} \in \mathcal{H}$ and $x \in \mathsf{X}$, $\|\hat{H}_{\boldsymbol{\eta}}(x)\| \leq L_{PH}^{(0)}, \|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x)\| \leq L_{PH}^{(0)}$, and

$$\sup_{x\in\mathsf{X}} \|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x)\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|, \ \forall \ (\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{H}^2.$$

(A7) It holds that $\sup_{\boldsymbol{\eta}\in\mathcal{H}, x\in\mathsf{X}} \|H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\| \leq \sigma$.

- **(A5)** refers to the existence of solution to Poisson equation.
- **(A6)** requires smoothness of $\hat{H}_{\boldsymbol{\eta}}(x) \Leftarrow$ satisfied if the $P_{\boldsymbol{\eta}}$, $H_{\boldsymbol{\eta}}(X)$ are smooth *w.r.t.* $\boldsymbol{\eta}$ + the Markov chain is geometrically ergodic.
- *Remark*: **(A7)** requires the update is *uniformly bounded* for all $x \in \mathsf{X}$. In fact, **(A5)–(A7)** can not imply **(A4)**, nor vice versa.

Case 2: State-dependent Markov Noise

Theorem 4

Let A1–A3, A5–A7 hold. Suppose that the step sizes satisfy

$$\gamma_{n+1} \leq \gamma_n, \ \gamma_n \leq a\gamma_{n+1}, \ \gamma_n - \gamma_{n+1} \leq a'\gamma_n^2, \ \gamma_1 \leq 0.5\big(c_1(L + C_h)\big)^{-1} ,$$

for $a, a' > 0$ *and all* $n \geq 0$*. Let* $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$,

$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] \leq \frac{2c_1\big(V_{0,n} + C_{0,n} + \big(\sigma^2 L + C_\gamma\big)\sum_{k=0}^{n}\gamma_{k+1}^2\big)}{\sum_{k=0}^{n}\gamma_{k+1}} + 2c_0 ,$$

- $C_h := \big(L_{PH}^{(1)}(d_0 + \frac{d_1}{2}(a+1) + ad_1\sigma) + L_{PH}^{(0)}\big(L + d_1\{1 + a'\}\big)\big)$.
- $C_\gamma := L_{PH}^{(1)}(d_0 + d_0\sigma + d_1\sigma) + LL_{PH}^{(0)}(1 + \sigma)$.
- $C_{0,n} := L_{PH}^{(0)}\big((1 + d_0)(\gamma_1 - \gamma_{n+1}) + d_0(\gamma_1 + \gamma_{n+1}) + 2d_1\big)$.

Case 2: State-dependent Markov Noise

Theorem 4

Let A1–A3, A5–A7 hold. Suppose that the step sizes satisfy

$$\gamma_{n+1} \leq \gamma_n, \; \gamma_n \leq a\gamma_{n+1}, \; \gamma_n - \gamma_{n+1} \leq a'\gamma_n^2, \; \gamma_1 \leq 0.5\big(c_1(L + C_h)\big)^{-1},$$

for $a, a' > 0$ *and all* $n \geq 0$. *Let* $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$,

$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] \leq \frac{2c_1\big(V_{0,n} + C_{0,n} + \big(\sigma^2 L + C_\gamma\big)\sum_{k=0}^{n}\gamma_{k+1}^2\big)}{\sum_{k=0}^{n}\gamma_{k+1}} + 2c_0 ,$$

- If $\gamma_k = (2c_1 L(1 + C_h)\sqrt{k})^{-1}$, then $\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] = \mathcal{O}(c_0 + \log n/\sqrt{n})$ as in our case 1 with Martingale noise.
- Key idea to the proof is to use the Poisson equation assumption **(A5)** [see Lemma 3], which is new to the SA analysis.

Latent Data Model

- **Goal:** Consider a stream of i.i.d. observations $\{Y_n\}_{n\geq 1}$, $Y_n \sim \pi$ and fit a parametric family $\{g(y;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$.
- Augment Y with *latent variable* $Z \Rightarrow$ complete data: $X = (Y, Z)$.
- **Exponential Family Distribution:** the complete data distribution:

$$f(x;\boldsymbol{\theta}) = h(x)\exp\left(\langle S(x)\,|\,\phi(\boldsymbol{\theta})\rangle - \psi(\boldsymbol{\theta})\right) ,$$

  where $S : \mathsf{X} \to \mathsf{S}$ is the sufficient statistics and $\mathsf{S} \subset \mathbb{R}^m$. We have $g(y;\boldsymbol{\theta}) = \int_{\mathsf{Z}} f(x;\boldsymbol{\theta})\mu(dz)$.

Two Important Operations

- **Expectation**: Given $\boldsymbol{\theta} \in \Theta$ and an observation y, the conditional expected sufficient statistics is

$$\bar{\boldsymbol{s}}(y; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[S(X) \,|\, Y = y \right]$$

- **Maximization**: Given a sufficient statistics $\boldsymbol{s} \in \mathsf{S}$, we can estimate $\boldsymbol{\theta}$ by maximizing the regularized log-likelihood

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) := \underset{\boldsymbol{\theta} \in \Theta}{\arg\max} \; \left\{ \langle \boldsymbol{s} \,|\, \phi(\boldsymbol{\theta}) \rangle - \psi(\boldsymbol{\theta}) - \mathsf{R}(\boldsymbol{\theta}) \right\}$$

where $\mathsf{R}(\boldsymbol{\theta})$ is a (strongly convex) regularization function.

Regularized Online EM (ro-EM)

- As $\{Y_n\}_{n\geq 1}$ arrive in a streaming fashion, the ro-EM method (modified from (Cappé and Moulines, 2009)) does:

$$\begin{aligned} \text{E-step:} \quad & \hat{\boldsymbol{s}}_{n+1} = \hat{\boldsymbol{s}}_n + \gamma_{n+1}\big\{\overline{\boldsymbol{s}}(Y_{n+1}; \hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{s}}_n\big\}, \\ \text{M-step:} \quad & \hat{\boldsymbol{\theta}}_{n+1} = \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_{n+1}). \end{aligned}$$

- Let us interpret **E-step** as an SA update (27) with drift term

$$H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) = \hat{\boldsymbol{s}}_n - \overline{\boldsymbol{s}}(Y_{n+1}; \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_n)) \ ,$$

whose mean field is given by $h(\hat{\boldsymbol{s}}_n) = \hat{\boldsymbol{s}}_n - \mathbb{E}_\pi\big[\overline{\boldsymbol{s}}(Y_{n+1}; \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_n))\big]$

- What should be the Lyapunov function? We use the KL divergence

$$V(\boldsymbol{s}) := \mathbb{E}_\pi\Big[\log \frac{\pi(Y)}{g(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))}\Big] + \mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})).$$

Special Case: Gaussian Mixture Model

Goal: fit $\{Y_n\}_{n\geq 1}$ in an GMM with $\boldsymbol{\theta} = (\{\omega_m\}_{m=1}^{M-1}, \{\mu_m\}_{m=1}^{M})$. Consider:

$$g(y;\boldsymbol{\theta}) \propto \left(1 - \sum_{m=1}^{M-1} \omega_m\right) \exp\left(-\frac{(y-\mu_M)^2}{2}\right) + \sum_{m=1}^{M-1} \omega_m \exp\left(-\frac{(y-\mu_m)^2}{2}\right),$$

and the regularizer (with $\epsilon > 0$)

$$\mathrm{R}(\boldsymbol{\theta}) = \epsilon \textstyle\sum_{m=1}^{M} \left\{\mu_m^2/2 - \log(\omega_m)\right\} - \epsilon \log\left(1 - \textstyle\sum_{m=1}^{M-1} \omega_m\right) .$$

Consider the assumption:

- **(A9)** The samples Y_n are i.i.d. and $|Y_n| \leq \overline{Y}$ for any $n \geq 0$.
- It can be verified that if **(A9)** holds, then **(A1)**, **(A3)**, **(A4)** are satisfied [cf. Proposition 3, 4].

Convergence Analysis

Corollary 1

Under A9 and set $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$*. The ro-EM method for GMM finds* $\hat{\boldsymbol{s}}_N$ *such that*

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}_N)\|^2] = \mathcal{O}(\log n/\sqrt{n})$$

The expectation is taken w.r.t. N *and the observation law* π*.*

- First *explicit non-asymptotic* rate given for online EM method.
- Note that rigorous convergence proof for *global convergence* of (online) EM methods are rare.

Policy Gradient: Markov Decision Process

- Consider a Markov Decision Process (MDP) $(\mathrm{S}, \mathrm{A}, \mathrm{R}, \mathrm{P})$:
  - S is a finite set of spaces (state-space)
  - A is a finite set of action (action-space)
  - $\mathrm{R} : \mathrm{S} \times \mathrm{A} \to [0, \mathrm{R}_{\max}]$ is a reward function
  - P is the transition model, *i.e.,* given an action $a \in \mathrm{A}$, $\mathrm{P}^a = \{\mathrm{P}^a_{s,s'}\}$ is a matrix, $\mathrm{P}^a_{s,s'}$ is the probability of transiting from the sth state to the s'th state upon taking action a.
- $\{(S_t, A_t)\}_{t \geq 1}$ forms a Markov chain with the transition probability from (s, a) to (s', a') as:

$$Q_{\boldsymbol{\eta}}((s,a);(s',a')) := \Pi_{\boldsymbol{\eta}}(a';s')\,\mathrm{P}^a_{s,s'} \ .$$

- The parameter $\boldsymbol{\eta}$ controls the *conditional probability of taking action* a' given the state s'.

Policy Optimization Problem

- **Goal:** Find a policy $\boldsymbol{\eta}$ to maximize the average reward:

$$J(\boldsymbol{\eta}) := \sum_{s \in \mathsf{S}, a \in \mathsf{A}} \upsilon(s,a)\, \mathsf{R}(s,a)\ .$$

  where $\upsilon(s,a)$ is the invariant distribution of $\{(S_t, A_t)\}_{t \geq 1}$.

- What is the gradient of $J(\boldsymbol{\eta})$ *w.r.t.* $\boldsymbol{\eta}$?

$$\nabla J(\boldsymbol{\eta}) = \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\eta}} \Big[\mathsf{R}(S_T, A_T) \sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\eta}}(A_{T-i}; S_{T-i}) \Big].$$

- REINFORCE algorithm (Williams, 1992) uses the sample average approximation. Let $M \gg 1$, $T \gg 1$,

$$\nabla J(\boldsymbol{\eta}) \approx (1/M) \sum_{m=1}^{M} \Big\{ \mathsf{R}(S_T^m, A_T^m) \sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\eta}}(A_{T-i}^m; S_{T-i}^m) \Big\}$$

  where $(S_1^m, A_1^m, \ldots, S_T^m, A_T^m) \sim \Pi_{\boldsymbol{\eta}}$ are drawn from a *roll-out* for each $m \Longrightarrow$ *needs many samples and $\boldsymbol{\eta}$ to be static.*

Policy Optimization Problem

- **Goal:** Find a policy $\boldsymbol{\eta}$ to maximize the average reward:

$$J(\boldsymbol{\eta}) := \sum_{s \in \mathsf{S}, a \in \mathsf{A}} \upsilon(s,a)\, \mathsf{R}(s,a)\ .$$

where $\upsilon(s,a)$ is the invariant distribution of $\{(S_t, A_t)\}_{t \geq 1}$.

- What is the gradient of $J(\boldsymbol{\eta})$ *w.r.t.* $\boldsymbol{\eta}$?

$$\nabla J(\boldsymbol{\eta}) = \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\eta}} \big[\mathsf{R}(S_T, A_T) \sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\eta}}(A_{T-i}; S_{T-i}) \big].$$

- We use a *biased* estimate of $\nabla J(\boldsymbol{\eta})$. Let $\lambda \in [0,1)$ and $T \gg 1$, we have (Baxter and Bartlett, 2001)

$$\nabla J(\boldsymbol{\eta}) \approx \widehat{\nabla}_T J(\boldsymbol{\eta}) := \mathsf{R}(S_T, A_T) \sum_{i=0}^{T-1} \lambda^i \nabla \log \Pi_{\boldsymbol{\eta}}(A_{T-i}; S_{T-i}),$$

where $(S_1, A_1, \dots, S_T, A_T) \sim \Pi_{\boldsymbol{\eta}}$.

Online Policy Gradient (PG)

- We update the policy on-the-fly with an online policy gradient update (Baxter and Bartlett, 2001; Tadić and Doucet, 2017):

$$G_{n+1} = \lambda G_n + \nabla \log \Pi_{\boldsymbol{\eta}_n}(A_{n+1}; S_{n+1}) \,, \tag{29a}$$

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_n + \gamma_{n+1} G_{n+1} \, \mathrm{R}(S_{n+1}, A_{n+1}) \,. \tag{29b}$$

- We can interpret (29b) as an SA step with the drift term:

$$H_{\boldsymbol{\eta}_n}(X_{n+1}) = G_{n+1} \, \mathrm{R}(S_{n+1}, A_{n+1})$$

- Let the joint state be $X_n = (S_n, A_n, G_n) \in \mathsf{S} \times \mathsf{A} \times \mathbb{R}^d$. We observe that $\{X_n\}_{n \geq 1}$ also forms a Markov chain. In particular,

$$h(\boldsymbol{\eta}) = \lim_{T \to \infty} \mathbb{E}_{\tau_T \sim \Pi_{\boldsymbol{\eta}},\ S_1 \sim \overline{\Pi}_{\boldsymbol{\eta}}} \big[\widehat{\nabla}_T J(\boldsymbol{\eta}) \big].$$

Convergence Analysis

- Focus on an exponential family policy (or soft-max):

$$\Pi_{\boldsymbol{\eta}}(a;s) = \left\{ \sum_{a' \in \mathsf{A}} \exp\left(\langle \boldsymbol{\eta} \,|\, \boldsymbol{x}(s,a') - \boldsymbol{x}(s,a) \rangle \right) \right\}^{-1} .$$

- **(A10)** For any $s \in \mathsf{S}$, we have $\|\boldsymbol{x}(s,a)\| \leq \overline{b}$.
- **(A11)** For any $\boldsymbol{\eta} \in \mathcal{H}$, $\{S_t, A_t\}_{t \geq 1}$ is geometrically ergodic with $\|\boldsymbol{Q}_{\boldsymbol{\eta}}^n - \mathbf{1}(\boldsymbol{\upsilon}_{\boldsymbol{\eta}})^\top\| \leq \rho^n K_R$. The invariant distribution $\boldsymbol{\upsilon}_{\boldsymbol{\eta}}$ and its Jacobian $\mathsf{J}^{\boldsymbol{\eta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\eta}}}(\boldsymbol{\eta})$ are Lipschitz continuous

$$\|\boldsymbol{\upsilon}_{\boldsymbol{\eta}} - \boldsymbol{\upsilon}_{\boldsymbol{\eta}'}\| \leq L_Q \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|, \ \| \mathsf{J}^{\boldsymbol{\eta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\eta}}}(\boldsymbol{\eta}) - \mathsf{J}^{\boldsymbol{\eta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\eta}}}(\boldsymbol{\eta}')\| \leq L_\upsilon \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|.$$

- Under **(A10)**, **(A11)**, the function $J(\boldsymbol{\eta})$ is $\mathsf{R}_{\max} |\mathsf{S}||\mathsf{A}|$-smooth,

$$(1-\lambda)^2 \Gamma^2 + 2 \langle \nabla J(\boldsymbol{\eta}) \,|\, h(\boldsymbol{\eta}) \rangle \geq \|h(\boldsymbol{\eta})\|^2,$$

where $\Gamma := 2\overline{b}\, \mathsf{R}_{\max} K_R \frac{1}{(1-\rho)^2}$. Other required assumptions are satisfied too [cf. Proposition 5-7].

Convergence Analysis (cont'd)

Corollary 2

Under A10, A11 and set $\gamma_k = (2c_1 L(1 + C_h)\sqrt{k})^{-1}$. *For any* $n \in \mathbb{N}$, *the policy gradient algorithm* (29) *finds a policy that*

$$\mathbb{E}\big[\|\nabla J(\boldsymbol{\eta}_N)\|^2\big] = \mathcal{O}\Big((1-\lambda)^2\Gamma^2 + c(\lambda)\log n/\sqrt{n}\Big), \tag{30}$$

where $c(\lambda) = \mathcal{O}(\frac{1}{1-\lambda})$. *Expectation is taken* w.r.t. N *and* (A_n, S_n).

- It shows the *first convergence rate* for the online PG method.
- Our result shows the *variance-bias trade-off* with $\lambda \in (0, 1)$.
- While setting $\lambda \to 1$ reduces the bias, but it decreases the convergence rate with $c(\lambda)$.

Sketches of Proofs

To simplify the notations next, we assume that $c_0 = 0, c_1 = 1$ and apply the following bound

Lemma 1. Let A1, A3 hold. With appropriate step size, it holds:

$$\sum_{k=0}^{n} (\gamma_{k+1}/2) \|h(\boldsymbol{\eta}_k)\|^2 \\ \leq V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1}) + L \sum_{k=1}^{n+1} \gamma_k^2 \|\boldsymbol{e}_k\|^2 - \sum_{k=1}^{n+1} \gamma_k \left\langle \nabla V(\boldsymbol{\eta}_{k+1}) \,|\, \boldsymbol{e}_k \right\rangle .$$

- **Case 1**: $\{\boldsymbol{e}_n\}_{n \geq 1}$ is Martingale $\Longrightarrow \mathbb{E}[\left\langle \nabla V(\boldsymbol{\eta}_{k+1}) \,|\, \boldsymbol{e}_k \right\rangle | \mathcal{F}_k] = 0$, taking the *total expectation* on both sides of the lemma suffices.
- **Case 2**: $\{\boldsymbol{e}_n\}_{n \geq 1}$ isn't Martingale $\Longrightarrow \mathbb{E}[\left\langle \nabla V(\boldsymbol{\eta}_{k+1}) \,|\, \boldsymbol{e}_k \right\rangle | \mathcal{F}_k] \neq 0$.
- In the latter case, we control the sum $\mathbb{E}[\sum_{k=1}^{n+1} \gamma_k \left\langle \nabla V(\boldsymbol{\eta}_{k+1}) \,|\, \boldsymbol{e}_k \right\rangle]$ using the Poisson's equation.

Consider the following bound

Lemma 2. Let A1–A3, A5–A7 hold. With appropriate step size, it holds:

$$\mathbb{E}\Big[-\sum_{k=0}^{n} \gamma_{k+1} \left\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \boldsymbol{e}_{k+1} \right\rangle \Big] \\ \leq C_h \sum_{k=0}^{n} \gamma_{k+1}^2 \mathbb{E}[\|h(\boldsymbol{\eta}_k)\|^2] + C_\gamma \sum_{k=0}^{n} \gamma_{k+1}^2 + C_{0,n},$$

- **Proof Idea**: note that $\boldsymbol{e}_{k+1} = H_{\boldsymbol{\eta}_k}(X_{k+1}) - h(\boldsymbol{\eta}_k)$, using **(A5)**, we have the decomposition

$$\sum_{k=0}^{n} \Big\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \hat{H}_{\boldsymbol{\eta}_k}(X_{k+1}) - P_{\boldsymbol{\eta}_k} \hat{H}_{\boldsymbol{\eta}_k}(X_{k+1}) \Big\rangle \equiv A_1 + A_2 + A_3 + A_4 + A_5$$

$$A_1 = \sum_{k=1}^{n} \Big\langle \nabla V(\boldsymbol{\eta}_k) \,|\, \hat{H}_{\boldsymbol{\eta}_k}(X_{k+1}) - P_{\boldsymbol{\eta}_k} \hat{H}_{\boldsymbol{\eta}_k}(X_k) \Big\rangle$$

$$A_2 = \sum_{k=1}^{n} \Big\langle \nabla V(\boldsymbol{\eta}_k) \,|\, P_{\boldsymbol{\eta}_k} \hat{H}_{\boldsymbol{\eta}_k}(X_k) - P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k) \Big\rangle$$

$$A_3 = \sum_{k=1}^{n} \Big\langle \nabla V(\boldsymbol{\eta}_k) - \nabla V(\boldsymbol{\eta}_{k-1}) \,|\, P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k) \Big\rangle$$

$$A_4 = \sum_{k=1}^{n} \Big\langle \nabla V(\boldsymbol{\eta}_k) \,|\, P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\eta}_{k-1}}(X_k) \Big\rangle$$

$$A_5 = \gamma_1 \Big\langle \nabla V(\boldsymbol{\eta}_0) \,|\, \hat{H}_{\boldsymbol{\eta}_0}(X_1) \Big\rangle - \gamma_{n+1} \Big\langle \nabla V(\boldsymbol{\eta}_n) \,|\, P_{\boldsymbol{\eta}_n} \hat{H}_{\boldsymbol{\eta}_n}(X_{n+1}) \Big\rangle$$

- We can bound A_1 by observing that the inner product is a Martingale, bound A_2 using **(A6)**, **(A7)**, etc..

3. Conclusion

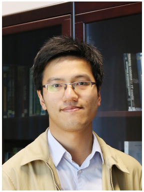
Take-aways

- We derived incremental and online methods for the optimization problem in machine learning.
    - For either ERM or Expected risk problems
    - When the objective function is a likelihood or not
    - For latent data models
- We conducted finite-time analysis of these methods for nonconvex loss functions and non necessarily gradient methods.
- Applications to several models of interest in machine learning with rigorous verification of the assumptions.

Perspectives

- Incremental algorithms: choice of the indices at each iteration. Optimal sampling strategies: (Roux et al., 2012) or (Horváth and Richtárik, 2018).
- Optimal mini-batch size of stochastic and incremental algorithms. See (Gower et al., 2019) (*variance-cost* trade off).
- Interplay between the Monte Carlo batch and the mini-batch of indices drawn at each iteration (*bias-variance* trade off).
- Complexity of $\mathcal{O}(n/\epsilon)$ was found for the MISO method. $\mathcal{O}(n^{2/3}/\epsilon)$ for quadratic surrogates in (Qian et al., 2019).
- Storage and Computation:
  - When data is big, need to store a lot for SAG, SAGA (less for SVRG).
  - Distributed first-order optimization procedures. Parallel or asynchronous method, or even centralized or decentralized architecture.





Thank you! Questions?

4. References

Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci*, 2:pp. 183.

Benveniste, A., Priouret, P., and Métivier, M. (1990). *Adaptive Algorithms and Stochastic Approximation*.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Cappé, O. and Moulines, E. (2009). On-line Expectation Maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

Dalal, G., Szorenyi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*.

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. (2012). Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578.

Fletcher, R., Gould, N. I., Leyffer, S., Toint, P. L., and Wächter, A. (2002). Global convergence of a trust-region sqp-filter algorithm for general nonlinear programming. *SIAM Journal on Optimization*, 13(3):635–659.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: general analysis and improved rates. *CoRR*, abs/1901.09401.

Horváth, S. and Richtárik, P. (2018). Nonconvex variance reduced optimization with arbitrary sampling. *arXiv preprint arXiv:1809.04146*.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019a). Non-asymptotic analysis of biased stochastic approximation scheme. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1944–1974, Phoenix, USA. PMLR.

Karimi, B., Wai, H.-T., and Moulines, E. (2019b). A doubly stochastic surrogate optimization scheme for non-convex finite-sum problems. *Submitted paper*.

Karimi, B., Wai, H.-T., Moulines, E., and Lavielle, M. (2019c). On the global convergence of (fast) incremental expectation maximization methods. *Submitted paper*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Mairal, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855.

McLachlan, G. and Krishnan, T. (2007). The EM Algorithm and Extensions.

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459.

Neal, R. and Hinton, G. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition.

Paisley, J., Blei, D., and Jordan, M. (2012). Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress.

Qian, X., Sailanbayev, A., Mishchenko, K., and Richtárik, P. (2019). Miso is making a comeback with better proofs and rates. *arXiv preprint arXiv:1906.01474*.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. (2016). Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Roux, N. L., Schmidt, M., and Bach, F. R. (2012). A stochastic gradient method with an exponential convergence _rate for finite training sets. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc.

Sun, T., Sun, Y., and Yin, W. (2018). On Markov chain gradient descent. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9918–9927. Curran Associates, Inc.

Tadić, V. B. and Doucet, A. (2017). Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1:pp. 1.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.