Accelerated Algorithms for a Bi-level Fixed-Point Learning Problem

Mouad Souissi[1], Alexandre Reiffers-Masson[1*]

[1]IMT Atlantique, Lab-STICC, 29200 Brest, France

[*]Corresponding author: firstname.lastname@imt-atlantique.fr

Background

Stochastic optimization methods play a central role in large-scale learning and statistical estimation. When the objective is expressed as a finite sum of loss functions:

$$f(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(x_i(\theta), z_i)$$

Standard stochastic approximation (SA) methods, such as Stochastic Gradient Descent (SGD) and Polyak–Ruppert Averaging (PR-SA), approximate the true gradient by sampling a subset of data points. However, uniform sampling induces high variance in the gradient estimator, leading to slow convergence [1] [3].

Importance Sampling (IS) provides a principled mechanism for **variance reduction** by reweighting samples according to their contribution to the estimator's variance. The recent paper *"Stochastic Optimization with Optimal Importance Sampling"* (Aolaritei & al, 2025) proposes a unified stochastic approximation framework where both the model parameters and the sampling distribution are jointly optimized to minimize the asymptotic variance of the iterates [1].

Our work explores how this **optimal importance sampling principle** can be adapted to a fixed-point learning problem, where each sample's contribution depends on an implicit function of the parameters. In this setting, gradients are defined through fixed-point equations, and classical assumptions such as independence or explicit gradient availability no longer hold directly.

Aim

The project aims to adapt the optimal importance sampling methodology to a stochastic optimization problem with implicit, fixed-point structures. Specifically, we seek to:

1. Extend the theoretical framework of importance sampling to stochastic gradients derived from fixed-point equations.
2. Establish the **asymptotic normality** of the iterates under the Martingale Central Limit Theorem (CLT) for dependent samples [2].
3. Design a **single-loop adaptive algorithm** that jointly optimizes the model parameters and the sampling distribution.
4. Demonstrate, through theoretical analysis, how **variance reduction and asymptotic optimality** can be achieved without prior knowledge of the optimal solution.

Method

Notation. We define the decision vector $\theta \in \mathbb{R}^K$. For each data point i, let $x_i(\theta) \in \mathbb{R}^{d_x}$ be the learned state and $u_i(\theta) \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$ be the learned control, where $\mathcal{U}$ is a closed, convex set. The dynamics are given by $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \times \mathcal{P} \to \mathbb{R}^{d_x}$, and the differentiable loss function is $L : \mathbb{R}^{d_x} \times \mathcal{Z} \to \mathbb{R}$. Finally, $\Gamma_{\mathcal{U}}$ denotes the Euclidean projection onto $\mathcal{U}$. We consider a **finite-sum stochastic optimization** problem where the gradient of each term depends on the implicit fixed points $(x_i(\theta), u_i(\theta))$:

$$x_i(\theta) = f(x_i(\theta), u_i(\theta), P_i), \quad u_i(\theta) = \Gamma_{\mathcal{U}}\left[u_i(\theta) - \alpha \nabla_u L\left(\sum_k \theta_k \hat{x}_i^{(k)}(\theta, u_i(\theta)), z_i\right)\right].$$

Under Lipschitz continuity and contraction assumptions, these fixed points are well-defined, allowing the use of **implicit differentiation** to compute gradients $G_i(\theta) = \nabla_\theta L(x_i(\theta), z_i)$.

We introduce a discrete IS distribution $q = (q_1, \ldots, q_N)$ over indices i, with likelihood ratio $\ell(i, q) = (1/N)/q_i$. The **optimal IS distribution** minimizing the variance of the stochastic gradient estimator is

$$q_i^* = \frac{\|G(\theta^*, i)\|}{\sum_j \|G(\theta^*, j)\|}.$$

This distribution oversamples gradients with large norms, thus reducing estimator variance.

Since q^* depends on the unknown optimum θ^*, we employ a **joint stochastic approximation** scheme, updating both θ and the IS parameters μ simultaneously:

$$\theta_{n+1} = \theta_n - \alpha_{n+1} G(\theta_n, i_n)\, \ell(i_n, \mu_n), \quad \mu_{n+1} = \mu_n - \alpha_{n+1} \|G(\theta_n, j_n)\|^2 \nabla_\mu \ell(j_n, \mu_n),$$

where $i_n \sim q(\mu_n)$ and j_n is uniformly sampled.

Preliminary Results

1. Asymptotic Convergence and Normality: By leveraging the Martingale CLT (Hall & Heyde, 1980), we expect to show that the averaged iterates satisfy

$$\sqrt{n}(\bar{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*),$$

with minimal asymptotic covariance

$$\Sigma^* = (\nabla^2 f(\theta^*))^{-1} \mathrm{Var}_{i \sim q^*}[G(\theta^*, i)] (\nabla^2 f(\theta^*))^{-1}.$$

2. Algorithmic Prototype and Empirical Validation: We implemented a prototype of the joint adaptive IS algorithm and compared it against classical SGD with uniform sampling on a simplified fixed-point learning model.

Figure 1 illustrates the performance of our prototype. The left panel demonstrates that the proposed Joint NDA (Adaptive IS) method achieves faster convergence and a lower terminal loss compared to the uniform baseline. The right panel tracks the entropy of the importance sampling distribution; its steady decrease from the maximum entropy (uniform) level confirms that the algorithm successfully adapts the sampling weights to focus on the most informative data points, thereby reducing variance.

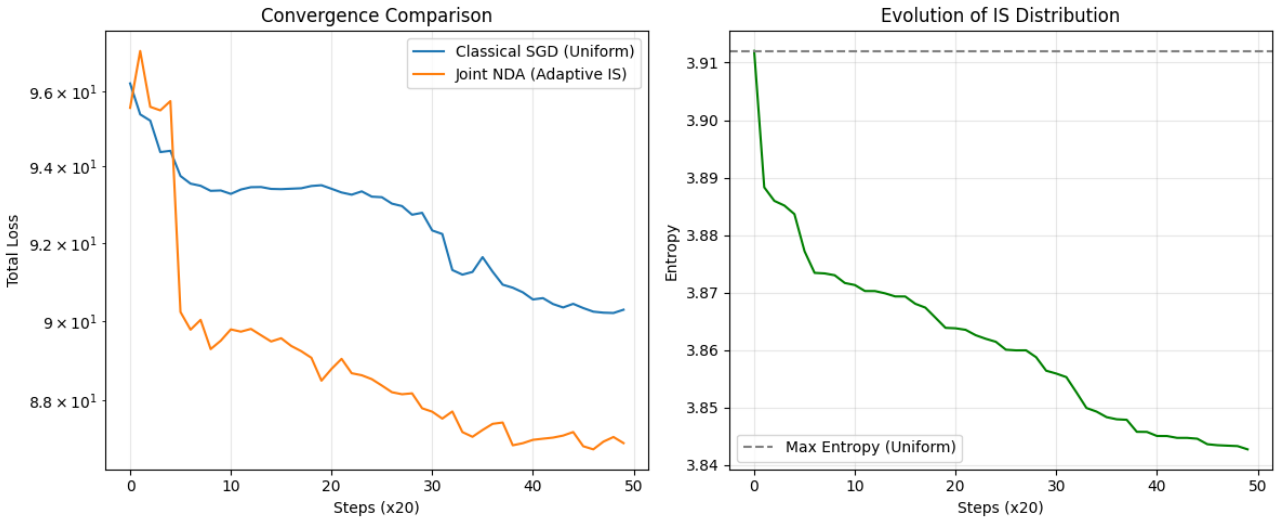

Figure 1: **Left:** Convergence comparison showing the total loss reduction of the Adaptive IS method vs. Classical SGD. **Right:** Evolution of the IS distribution entropy, showing deviation from the uniform distribution.

References

1. L. Aolaritei, B. P. Van Parys, H. Lam, and M. I. Jordan. (2025) *Stochastic Optimization with Optimal Importance Sampling. arXiv preprint arXiv:2504.03560*
2. Hall, P., & Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.
3. Polyak, B. T., & Juditsky, A. B. (1992). *Acceleration of stochastic approximation by averaging*. SIAM Journal on Control and Optimization, 30(4), 838–855.