Asymptotic Analysis of Conditioned Stochastic Gradient Descent

Rémi LELUC

Ecole Polytechnique, Institut Polytechnique de Paris, France

Joint Work with François Portier, paper

Accepted at *Transactions on Machine Learning Research, 2023*

We consider the following type of optimization problem:

$$\min_{x \in \mathbb{R}^d} \{F(x) = \mathbb{E}_\xi[f(x,\xi)]\}$$

∇F **hard to compute** (ERM) or **intractable** (AIS and RL)

Unbiased estimate in SGD

There is a **cheap** gradient generator $g(\cdot,\xi)$ s.t.

$$\forall x \in \mathbb{R}^d, \qquad \mathbb{E}_\xi[g(x,\xi)] = \nabla F(x)$$

$$(SGD) \quad x_{k+1} = x_k - \alpha_{k+1} g(x_k, \xi_{k+1})$$

- Reference book on Stochastic programming (Shapiro et al., 2014)
- Comparison with sample average approximation (Nemirovski et al., 2009)

Examples

Empirical Risk Minimization (ERM). Data $z_1, \ldots, z_N \in \mathcal{Z}$, loss $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, empirical risk $F(x) = N^{-1} \sum_{i=1}^{N} \ell(x, z_i) = \int \ell(x, z) P_N(dz)$

$$g(x, \xi) = \nabla_x \ell(x, \xi), \qquad \xi \sim P_N$$

Recursive estimates (at time t, a new variable is observed $z_t \sim P$)

Adaptive importance sampling (AIS). Target function f, parametric family of samplers $\{q_x \,:\, x \in \Theta\}$, objective $F(x) = -\int \log(q_x(y)) f(y) dy$

$$g(x, \xi) = -\nabla_x \log(q_x(\xi)) \frac{f(\xi)}{q_0(\xi)}, \qquad \xi \sim q_0.$$

Policy-gradient methods (RL). Algorithm REINFORCE uses a parameterized policy $\{\pi_x \,:\, x \in \Theta\}$ to optimize expected reward $F(x) = \mathbb{E}_{\xi \sim \pi_x}[\mathcal{R}(\xi)]$

$$g(x, \xi) = \mathcal{R}(\xi) \nabla_x \log \pi_x(\xi), \qquad \xi \sim \pi_x.$$

Robbins and Monro (1951)

$$(SGD) \quad x_{k+1} = x_k - \alpha_{k+1} g(x_k, \xi_{k+1})$$

Stochastic approximation literature

- **Almost sure convergence:** Robbins and Siegmund (1971) and Bertsekas and Tsitsiklis (2000)
- **Rates of convergence** and **central limit theorem:** Sacks (1958); Kushner and Huang (1979), a law of the iterated logarithm by Pelletier (1998a)
- Two different approaches for the asymptotic normality: **diffusion-based method** (Benaïm, 1999; Pelletier, 1998b; Gadat et al., 2018);**martingale tools** (Kushner and Clark, 1978; Delyon, 1996; Hall and Heyde, 2014); **Review:** (Lai et al., 2003)

ML point of view

- Review paper: Bottou et al. (2018),
- Non asymptotic bounds (Moulines and Bach, 2011)

Conditioned-SGD (CSGD)

$$(CSGD) \quad x_{k+1} = x_k - \alpha_{k+1} C_k g(x_k, \xi_{k+1})$$

Optimal choice: $C_k \simeq \nabla^2 F(x^*)^{-1}$ (Newton's method)

Methods

- Approximation of $\nabla^2 F(x^*)^{-1}$ based on Taylor expansion (Agarwal et al., 2016).
- Ricatti's formula in *logistic regression* (Bercu et al., 2020); generalized in Boyer and Godichon-Baggioni (2020).
- **Fisher information** matrix (Amari (1998); Kakade (2002)).
- BFGS approximation in ML literature (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970; Byrd et al., 2011; Moritz et al., 2016).

Questions raised

- What condition on C_k to ensure convergence ? Asymptotic normality ?
- What conditioning matrix C_k should we choose ?

$\rightarrow$ The optimal choice according to the asymptotic variance is the inverse of the Hessian matrix $C_k = \nabla F(x^\star)^{-1}$

- Is this optimal variance achieved by a feasible algorithm ?

$\rightarrow$ We show that the answer is positive under mild conditions on the matrix $C_k = \nabla F(x^\star)^{-1}$

Some answers

- **SA literature:** Venter (1967); Fabian et al. (1973); Nevelson and Hasminskii (1976); Ruppert et al. (1985); Wei et al. (1987); Spall (2000)
- The CLT given in Pelletier (1998b) requires that $\|C_k - C^*\| \ll \|x_k - x^*\|$
- Boyer and Godichon-Baggioni (2020) works for convex functions and requires $\|C_k - C^*\| = O(n^{-s})$, $s > 0$.

Contributions

$$(CSGD) \quad x_{k+1} = x_k - \alpha_{k+1} C_k g(x_k, \xi_{k+1})$$

Online and nonconvex optimization

- L-smoothness, growth conditions
- the gradient policy is allowed to change in time

A continuity result for CSGD's weak limit

- If $C_k \to C^*$ a.s., then CSGD with $C_k \simeq$ CSGD with C^*
- We give an example where efficiency is reached

Stochastic equicontinuity of the empirical process: if $(X_i)_{i \geq 1}$ is **iid** and $\{f_\eta\}$ with **small complexity**, then $\int (f_{\hat{\eta}_n}(x) - f_{\eta_0}(x))^2 P(dx) = o_P(1)$ implies that

$$\mathbb{G}_n\{f_{\hat{\eta}_n} - f_{\eta_0}\} = o_P(1)$$

(**in words**: estimating η_0 has no effect at the limit)

Mathematical Background

Goal

Find the minimizer $x^\star \in \mathbb{R}^d$ of a function $F : \mathbb{R}^d \to \mathbb{R}$,

$$x^\star = \arg\min_{x \in \mathbb{R}^d} F(x).$$

No convexity required on F

- F is L-smooth, coercive and the equation $\nabla F(x) = 0$ has unique solution $x^\star$.
- $H = \nabla^2 F(x^*) \succ 0$ and $\nabla^2 F$ is continuous at x^*

Asymptotics of SGD

SGD policy (Gower et al., 2019)

$$(SGD) \qquad x_{k+1} = x_k - \alpha_{k+1} g(x_k, \xi_{k+1}),$$

with $\forall x \in \mathbb{R}^d, \forall k \in \mathbb{N}$

- $\mathbb{E}[g(x, \xi_{k+1}) | \mathcal{F}_k] = \nabla F(x)$
- $\mathbb{E}\left[\|g(x, \xi_{k+1})\|^2 | \mathcal{F}_k\right] \leq 2\mathcal{L}\left(F(x) - F(x^\star)\right) + \sigma^2$

Robbins-Monro condition

$(\alpha_k)_{k\geq 1} \downarrow 0$ s.t. $\sum_{k\geq 1} \alpha_k = +\infty$ $\sum_{k\geq 1} \alpha_k^2 < +\infty$

- In practice $\alpha_k = \alpha k^{-\beta}$, $\beta \in (1/2, 1]$

Theorem (Almost sure convergence)

$$x_k \to x^\star \qquad a.s.$$

Additional assumptions

- Liapounov condition and limiting variance Γ on the martingale increments
- $\alpha_k = \alpha k^{-\beta}$, $\beta \in (1/2, 1]$
- $(H - \kappa I) \succ 0$ with $\kappa = 1_{\{\beta=1\}} 1/2\alpha$

Theorem (Weak convergence Pelletier (1998b))

The SGD rule satisfies

$$\frac{1}{\sqrt{\alpha_k}}(x_k - x^\star) \rightsquigarrow \mathcal{N}(0, \Sigma), \qquad \text{as } k \to \infty$$

where Σ satisfies the following Lyapunov equation

$$(H - \kappa I)\Sigma + \Sigma(H^T - \kappa I) = \Gamma.$$

- Fastest rate for $\beta = 1$ and recover the classical $1/\sqrt{k}$-rate.
- α large enough to ensure $H - I/(2\alpha) \succ 0$ but small enough so that $\alpha\Sigma$ small.

Variance optimality when $\beta = 1$ via Conditioning

Replace the scalar gain α by a conditioning matrix $C \in \mathbb{R}^{d\times d}$

$$x_{k+1} = x_k - \left(\frac{C}{k+1}\right) g(x_k, \xi_{k+1}).$$

with $CH - \kappa I \succ 0$.

Theorem (Deterministic Conditioning)

The sequence $(x_k)_{k\geq 0}$ (given above) satisfies

$$\sqrt{k}(x_k - x^\star) \rightsquigarrow \mathcal{N}(0, \Sigma_C)$$

where Σ_C is given by the Liapounov equation

$$\left(CH - \frac{I}{2}\right)\Sigma_C + \Sigma_C\left((CH)^T - \frac{I}{2}\right) = C\Gamma C^T.$$

Deterministic Conditioning

What conditioning matrix C should we choose ?

Optimal Variance

The choice $C^\star = H^{-1}$ is optimal: $\Sigma_{C^\star} \leq \Sigma_C \ \forall C$

- (Asymptotic efficiency) $\sqrt{k}(x_k - x^\star) \rightsquigarrow \mathcal{N}(0, \Sigma_{C^\star} = H^{-1}\Gamma H^{-1})$
- Averaging of standard SGD gives the same variance (Polyak and Juditsky, 1992)
- $C^\star$ is usually unknown ...

Conditioned-SGD

CSGD

$$(CSGD) \quad x_{k+1} = x_k - \alpha_{k+1} C_k g(x_k, \xi_{k+1})$$

with

$$\beta_k I_d \leq C_{k-1} \leq \gamma_k I_d$$

Extended Robbins-Monro

The sequences $(\alpha_k)_{k \geq 1}, (\beta_k)_{k \geq 1}, (\gamma_k)_{k \geq 1}$ are positive and satisfy

$$\sum_{k \geq 1} \alpha_k \beta_k = +\infty \qquad \sum_{k \geq 1} (\alpha_k \gamma_k)^2 < +\infty$$

- Note that $C_k = I_d$ recovers SGD with standard Robbins-Monro.

Theorem (Almost sure convergence)

The sequence of CSGD iterates satisfies $x_k \to x^\star \quad a.s.$

- At what speed $(x_k - x^\star)$ is bounded ? Asymptotic normality ?

CSGD

Mild assumption on the conditioning matrices

- $C_k \to C$ a.s.
- $(CH - \kappa I)$ positive definite with $\kappa = 1_{\{\beta=1\}} 1/2\alpha$

Theorem (Weak convergence)

The sequence of CSGD satisfies

$$\frac{1}{\sqrt{\alpha_k}}(x_k - x^\star) \rightsquigarrow \mathcal{N}(0, \Sigma_C), \qquad \text{as } k \to \infty,$$

where Σ_C satisfies:

$$(CH - \kappa I)\,\Sigma_C + \Sigma_C\left((CH)^T - \kappa I\right) = C\Gamma C^T.$$

- Continuity property (as if $C_k = C$)
- C should be close to the inverse of the Hessian $H = \nabla^2 F(x^\star)$.

Sketch of the proof

In a similar spirit as in Delyon (1996), the proof relies on the introduction of a linear stochastic algorithm based on the approximation

$$\nabla F(x_{k-1}) \simeq H(x_{k-1} - x^\star)$$

We consider the auxiliary iteration

$$\widetilde{\Delta}_k = \widetilde{\Delta}_{k-1} - \alpha_k K \widetilde{\Delta}_{k-1} - \alpha_k C_{k-1} w_k, \qquad k \geq 1$$

with $K = CH$ and $w_k = g(x_{k-1}, \xi_k) - \nabla f(x_{k-1})$. Then we show that

$$(x_k - x^\star) - \widetilde{\Delta}_k = o_{\mathbb{P}}(\sqrt{\alpha_k})$$

The analysis of $\widetilde{\Delta}_k / \sqrt{\alpha_k}$ is carried out with martingale tools.

An effective algorithm

Hessian generator

There exists Hessian generator $H(\cdot, \xi'_{k+1})$ such that

$$\forall k \geq 1,\, \forall x \qquad \mathbb{E}\left[H(x, \xi'_{k+1}) | \mathcal{F}_k\right] = \nabla^2 F(x).$$

Average past estimates with some weights

$$\Phi_k = \sum_{j=0}^{k} \nu_{j,k} H(x_j, \xi'_{j+1}),$$

where $\nu_{j,k} \propto \exp(-\eta \|x_j - x_k\|_1)$ is such that $\sum_{j=0}^{k} \nu_{j,k} = 1$.

Regularization

$$\forall k \in \mathbb{N}, \quad C_k = \left(\Phi_k + \gamma_{k+1}^{-1} I_d\right)^{-1}$$

Asymptotic Efficiency of CSGD

Proposition

If $H(x,\xi)$ *bounded and* $\gamma_k \to \infty$ *then* $C_k \to H^{-1}$ $a.s$

(proof: Freedman's inequality and the Cesaro Lemma)

Corollary (Asymptotic optimality)

Let $(x_k)_{k\geq 0}$ *be the CSGD iterates with* $\alpha_k = 1/k$ *and* C_k *given before. If* $\sum_{k\geq 1}(\gamma_k/k)^2 < \infty$, *we have*

$$\sqrt{k}(x_k - x^\star) \rightsquigarrow \mathcal{N}(0, H^{-1}\Gamma H^{-1}), \qquad \text{as } k \to \infty$$

- Asymptotic optimality is reached !
- Practical choice $\alpha_k = 1/k$: removes the assumption $2\alpha H \succ I$

Corollary (Asymptotic of excess risk)

$$k(F(x_k) - F(x^\star)) \rightsquigarrow \sum_{k=1}^{d} \lambda_k Z_k^2,$$

$(Z_1, \ldots, Z_d) \sim \mathcal{N}(0, I_d)$ *and* $(\lambda_k)_{k=1}^{d}$ *are eigenvalues of* $H^{-1/2}\Gamma H^{-1/2}$

Numerical Experiments: ERM

- We apply ERM to regularized **regression** and **classification** problems.

Ridge regression

Given a data matrix $X = (x_{i,j}) \in \mathbb{R}^{n \times p}$ with labels $y \in \mathbb{R}^n$ and a regularization parameter $\mu > 0$. Consider

$$\min_{\theta \in \mathbb{R}^d} F(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} x_{i,j}\theta_j)^2 + \frac{\mu}{2}\|\theta\|_2^2$$

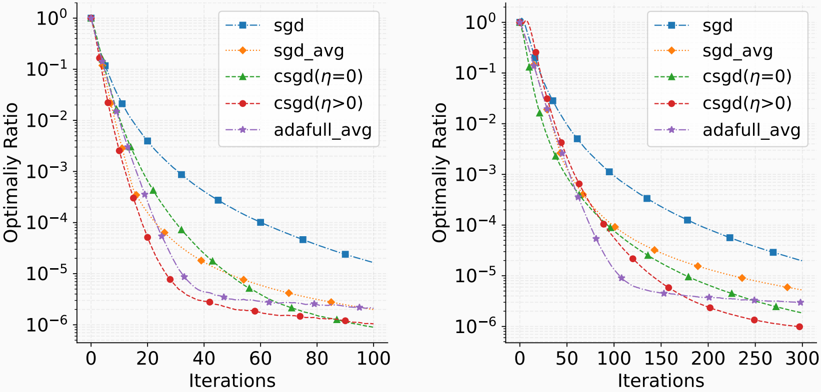
logistic regression

Given a data matrix $X = (x_{i,j}) \in \mathbb{R}^{n \times p}$ with labels $y \in \mathbb{R}^n$ and a regularization parameter $\mu > 0$. Consider

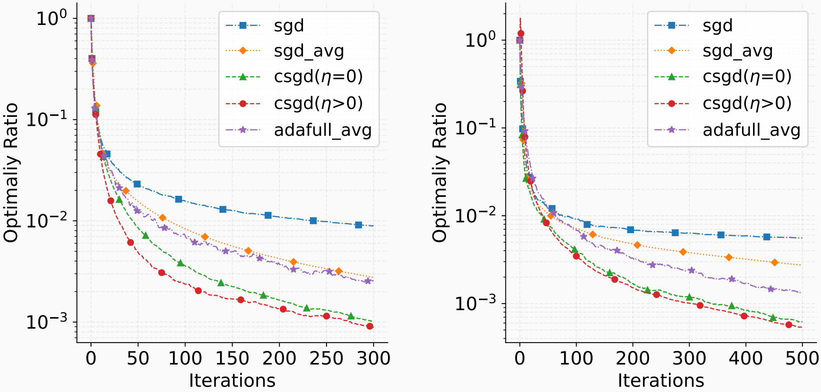
$$\min_{\theta \in \mathbb{R}^d} F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \sum_{j=1}^{d} x_{i,j}\theta_j)) + \mu\|\theta\|_2^2$$

- Setting $n = 10,000$, $d \in \{20, 100\}$, $\mu = 1/n$.

Numerical Experiments: synthetic $d = 20$ and $d = 100$



Numerical Experiments: Boston and Diabetes datasets



Conclusion

Contributions

- Almost sure convergence of CSGD in a **non convex setting**
- Asymptotic normality: **Equi-continuity property when** $C_k \to C$
- Definition of an algorithm that achieves **efficiency**

Applications

$\to$ When the Hessian is known exactly without noise
$\to$ Dynamical update of Hessian estimates (BFGS)
$\to$ Particular choice of diagonal conditioning matrix with weights: perform coordinate sampling, see our paper at *Journal of Machine Learning Research, 2022*

References

Agarwal, N., B. Bullins, and E. Hazan (2016). Second-order stochastic optimization in linear time. *stat 1050*, 15.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation 10*(2), 251–276.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pp. 1–68. Springer.

Bercu, B., A. Godichon, and B. Portier (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization 58*(1), 348–367.

Bertsekas, D. P. and J. N. Tsitsiklis (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization 10*(3), 627–642.

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

Boyer, C. and A. Godichon-Baggioni (2020). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *arXiv preprint arXiv:2011.09706*.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics 6*(1), 76–90.

Byrd, R. H., G. M. Chin, W. Neveitt, and J. Nocedal (2011). On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization 21*(3), 977–995.

Delyon, B. (1996). General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control 41*(9), 1245–1255.

Fabian, V. et al. (1973). Asymptotically efficient stochastic approximation; the rm case. *The Annals of Statistics 1*(3), 486–495.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal 13*(3), 317–322.

Gadat, S., F. Panloup, S. Saadane, et al. (2018). Stochastic heavy ball. *Electronic Journal of Statistics 12*(1), 461–529.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation 24*(109), 23–26.

Gower, R. M., N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik (2019). Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*.

Hall, P. and C. C. Heyde (2014). *Martingale limit theory and its application*. Academic press.

Kakade, S. M. (2002). A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538.

Kushner, H. J. and D. S. Clark (1978). Stochastic approximation methods for constrained and unconstrained systems.

Kushner, H. J. and H. Huang (1979). Rates of convergence for stochastic approximation type algorithms. *SIAM Journal on Control and Optimization 17*(5), 607–617.

Lai, T. L. et al. (2003). Stochastic approximation. *Annals of Statistics 31*(2), 391–406.

Moritz, P., R. Nishihara, and M. Jordan (2016). A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pp. 249–258. PMLR.

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

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization 19*(4), 1574–1609.

Nevelson, M. B. and R. Z. Hasminskii (1976). *Stochastic approximation and recursive estimation*, Volume 47. Amer Mathematical Society.

Pelletier, M. (1998a). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications 78*(2), 217–244.

Pelletier, M. (1998b). Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Annals of Applied Probability*, 10–44.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization 30*(4), 838–855.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Robbins, H. and D. Siegmund (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier.

Ruppert, D. et al. (1985). A newton-raphson version of the multivariate robbins-monro procedure. *Annals of Statistics 13*(1), 236–245.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics 29*(2), 373–405.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation 24*(111), 647–656.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

Spall, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE transactions on automatic control 45*(10), 1839–1853.

Venter, J. H. (1967). An extension of the robbins-monro procedure. *The Annals of Mathematical Statistics 38*(1), 181–190.

Wei, C. et al. (1987). Multivariate adaptive stochastic approximation. *The Annals of Statistics 15*(3), 1115–1130.