Monte Carlo Methods and Stochastic Approximation: Theory and Applications to Machine Learning

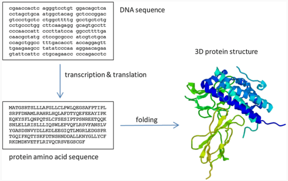
Rémi LELUC, *PhD defense*
LTCI, Télécom Paris, France

Jury:

BACH Francis	Examiner
BIANCHI Pascal	co-Supervisor
CARPENTIER Alexandra	Examiner
CHOPIN Nicolas	President
GADAT Sébastien	Reviewer
MERTIKOPOULOS Panayotis	Examiner
PORTIER François	Supervisor
ROBERT Christian	Reviewer

Motivation: Machine Learning recent advances

AlphaGo (2016)



AlphaFold (2018)

GPT-3/4(2020/2023)

Machine Learning goal

Learn (*integrate*/*optimize*) a prediction function

Motivation: need for integral and gradient estimators

Central Question 1: *Integration*

Computation of an *integral* through probabilistic objective $\mathcal{F}$

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi_\theta(x)}[f(x)] = \int_{\mathcal{X}} f(x)\pi_\theta(x)\mathrm{d}x. \tag{1}$$

Cost function f and input distribution $\pi_\theta(\cdot)$

Central Question 2: *Optimization*

Learn the optimal parameter $\theta^\star \in \arg\min_\theta \mathcal{F}(\theta)$ with the gradient

$$\mathcal{G} = \nabla_\theta \mathcal{F}(\theta) = \nabla_\theta \, \mathbb{E}_{\pi_\theta(x)}[f(x)]. \tag{2}$$

Main issue: intractability and computational cost

Reinforcement Learning[1].

Trajectory $\tau = (s_0, a_0, \ldots, s_{T-1}, a_{T-1})$ with policy π_θ and cumulative return $\mathcal{R}(\tau) = \sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$.

Objective $\mathcal{F}$ is an *expectation*

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi_\theta(\tau)}[\mathcal{R}(\tau)]$$

Optimal strategy $\pi_{\theta^\star}$ with $\theta^\star \in \arg\max \mathcal{F}(\theta)$

(2016) AlphaGo A.I. beats champion Lee Sedol in Go.

Rely on gradient-based *optimization* techniques with gradient

$$\mathcal{G} = \mathbb{E}_{\pi_\theta(\tau)}[\mathcal{R}(\tau)\nabla_\theta \log \pi_\theta(\tau)].$$

[1](Sutton and Barto, 2018): Reinforcement Learning: An introduction

Advantages of Random estimates

Easy and Practical

$\rightarrow$ Requires only three steps: sampling, evaluating, averaging

Randomness as a Strength

$\rightarrow$ Naturally escape local optima[2]

$\rightarrow$ Complete exploration of the search space

Large-Scale learning

$\rightarrow$ simple, scalable, parallelizable

$\rightarrow$ in supervised learning, deterministic gradient scales as $O(nd)$, stochastic version reduces to $O(d)$ operations

Theoretical justifications[3]

$\rightarrow$ deterministic methods $O(n^{-s/d})$

$\rightarrow$ optimal random procedure $O(n^{-1/2}n^{-s/d})$

[2](Gadat et al., 2018): Stochastic heavy ball

[3](Novak, 2016): Some results on the complexity of numerical integration

Outline for today

$$\textit{Integrate } \mathcal{F}(\theta) = \int_{\mathcal{X}} f(x)\pi_\theta(\mathrm{d}x) \to \textit{Optimize } \mathcal{F} \text{ with } \nabla\mathcal{F}$$

Part I: Monte Carlo Integration (approximate $\mathcal{F}(\theta)$)

Part II: Stochastic Optimization Methods (optimize $\mathcal{F}$)

Part I: Integration $\mathcal{F}$

Monte Carlo Integration, Variance Reduction

1. **R. Leluc**, F. Portier and J. Segers. *Control Variate Selection for Monte Carlo Integration.* (Leluc et al., 2021)
In *Statistics and Computing 31, 50*, pages1-27, 2021.

2. **R. Leluc**, F. Portier, J. Segers and A. Zhuman. *A Quadrature Rule combining Control Variates and Adaptive Importance Sampling.* (Leluc et al., 2022)
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

3. **R. Leluc**, F. Portier, J. Segers and A. Zhuman. *Speeding up Monte Carlo Integration: Nearest Neighbors as Control Variates. arXiv preprint*, 2023.

Monte Carlo integration

Underlying integration problem

Let $(\mathcal{X}, \mathcal{A}, \pi)$ be a probability space, $f : \mathcal{X} \to \mathbb{R}$ with $f \in L_2(\pi)$.

- **Goal:**

$$\pi(f) := \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x) = \mathbb{E}_\pi[f(X)].$$

- **Constraints:** f is unknown (black-box) or no approximation is sufficiently accurate, sampling from π may be hard.

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \pi$, naive Monte Carlo estimator $\hat{\alpha}_n^{\mathrm{mc}}(f)$ of $\pi(f)$ is

$$\hat{\alpha}_n^{\mathrm{mc}}(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tag{3}$$

Research Questions (Part I)

- How to reduce the variance of Monte Carlo estimates?
- How to sample from π?
- How to achieve optimal convergence rates?

Ref: Metropolis and Ulam (1949); Robert and Casella (1999); Evans and Swartz (2000); Glasserman (2004); Owen (2013); Novak (2016); Chopin and Gerber (2022)

Variance Reduction with Control Variates

Definition: Control Variates

Functions $h_1, \ldots, h_m \in L_2(\pi)$ with known integrals:

$$\forall 1 \leq j \leq m, \quad \mathbb{E}_\pi[h_j] = 0$$

$\to$ Stein control variates, families of orthogonal polynomials

- Let $h = (h_1, \ldots, h_m)^\top$, for any $\beta \in \mathbb{R}^m$, we have $\mathbb{E}_\pi[f - \beta^\top h] = \mathbb{E}_\pi[f]$ leading to the CV estimate of α, parameterized by β

CV-Monte Carlo

$$\alpha_n^{(\mathrm{cv})}(f, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - \beta^\top h(X_i)\right), \quad X_1, \ldots, X_n \sim \pi.$$

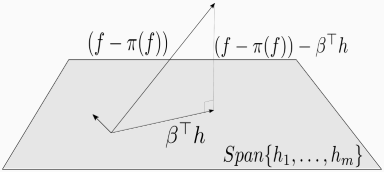
- What optimal choice for β^* ? Look at variance and define

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^m} \mathbb{E}_\pi \left[(f - \pi(f) - \beta^\top h)^2\right]$$

Integration with Linear regression

From integration to linear regression

The integral $\pi(f)$ appears as the intercept of a linear regression model with response f and explanatory variables $h_1, \ldots, h_m$,



L_2-orthogonal projection.

- The integral and oracle coefficient satisfy

$$(\pi(f), \beta^\star(f)) \in \operatorname*{arg\,min}_{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m} \pi[(f - \alpha - \beta^\top h)^2] \tag{4}$$

- Replacing the distribution π by the sample measure $\hat{\pi}_n$ gives the **Ordinary Least Squares** (OLS) estimate, $X_1, \ldots, X_n \sim \pi$

$$\left(\hat{\alpha}_n^{(\mathrm{cv})}, \hat{\beta}_n^{(\mathrm{cv})}\right) \in \operatorname*{arg\,min}_{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \left(f(X_i) - \alpha - \beta^\top h(X_i)\right)^2 \tag{5}$$

From Ordinary Least Squares Monte Carlo...

Limitations of OLSMC.

- (*Overfitting*) Too many variables or/and few samples (case $m >> n$)
- (*Collinearity*) Dependence among variables $\rightarrow$ very large coefficients

How to avoid those problems ?

Limitations of OLSMC.

- (*Overfitting*) Too many variables or/and few samples (case $m >> n$)
- (*Collinearity*) Dependence among variables $\rightarrow$ very large coefficients

How to avoid those problems ?

Bet on sparsity with **variable selection**!

Image generated by text-to-image A.I. midjourney with the command: "super-hero cowboy twirling his lasso in the air, comic-book style".

... to Lasso Monte-Carlo (LASSOMC/LSLASSO)

Control Variates estimates: OLS, LASSO, LSLASSO

$$\left(\hat{\alpha}_n^{\mathrm{ols}}(f), \hat{\beta}_n^{\mathrm{ols}}(f)\right) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \|f^{(n)} - \alpha\mathbb{1}_n - H\beta\|_2^2$$

$$\left(\hat{\alpha}_n^{\mathrm{lasso}}(f), \hat{\beta}_n^{\mathrm{lasso}}(f)\right) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \frac{1}{2n}\|f^{(n)} - \alpha\mathbb{1}_n - H\beta\|_2^2 + \lambda\|\beta\|_1$$

$$\left(\hat{\alpha}_n^{\mathrm{lslasso}}(f), \hat{\beta}_n^{\mathrm{lslasso}}(f)\right) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^{\hat{\ell}}}{\arg\min} \|f^{(n)} - \alpha\mathbb{1}_n - H_{\hat{S}}\beta\|_2^2$$

- **Active set** $S^\star = \{k : \beta_k^\star \neq 0\}$ and **sparsity level** $\ell^\star = Card(S^\star)$
- LSLASSOMC:

(1) $\hat{S} = \{k : \hat{\beta}_{N,k}^{\mathrm{lasso}}(f) \neq 0\}$ estimated **active set** with **LASSO**
(2) Solve subproblem **OLS** with selected control variates

Non-asymptotic Error Analysis

Assumptions: **sub-gaussian** residuals $\varepsilon = f - \pi(f) - \beta^{\star\top} h$ with factor τ.

Concentration inequalities

For $\delta \in (0,1)$ with probability at least $1-\delta$, for OLS, LASSO, LSLASSO

$$|\hat{\alpha}_n^{\mathrm{ols}}(f) - \pi(f)| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + C_1\sqrt{Bm\log(8m/\delta)}\frac{\tau}{n}$$

$$|\hat{\alpha}_n^{\mathrm{lasso}}(f) - \pi(f)| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + C_2(U_h^2/\gamma^\star)\ell^\star\log(8m/\delta)\frac{\tau}{n}$$

$$|\hat{\alpha}_n^{\mathrm{lslasso}}(f) - \pi(f)| \leq \sqrt{2\log(16/\delta)}\frac{\tau}{\sqrt{n}} + C_3\sqrt{B^\star\ell^\star\log(16\ell^\star/\delta)}\frac{\tau}{n}$$

$U_h = \max_{j=1,\dots,m} \|h_j\|_\infty$

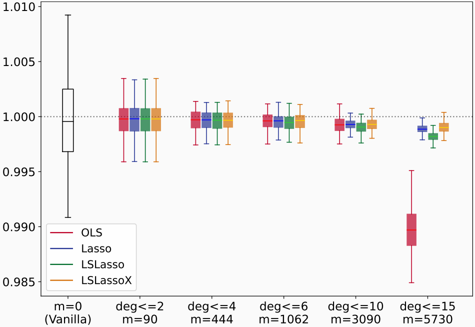
$G = \mathbb{E}_\pi[hh^\top], \gamma = \lambda_{\min}(G), \hbar = G^{-1/2}h$; $B = \sup_x \|\hbar(x)\|_2^2$

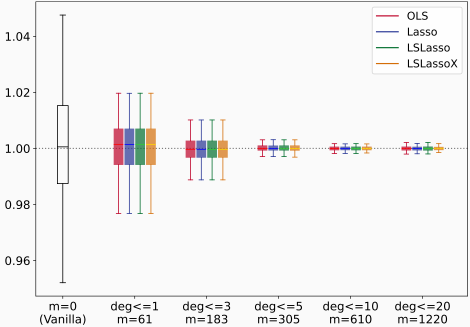
$G^\star, \gamma^\star, B^\star$ restricted on **active set**

Evidence Estimation in Bayesian Models

- Model likelihood $\ell(x|\theta)$ and prior distribution $\pi(\theta)$, compute evidence

$$Z = \int_{\Theta} \ell(x|\theta)\pi(\theta)\mathrm{d}\theta$$





Boxplots of Error Distribution for Capture ($d = 12$) and Sonar ($d = 61$) datasets[4], $n = 5000$; $N = 1000$, obtained over 100 replications.

[4] (Marzolin, 1988; Gorman and Sejnowski, 1988)

Monte Carlo Integration and Importance Sampling

GOAL:

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x$$

Can we sample from target distribution π ?

Monte Carlo Integration and Importance Sampling

GOAL:

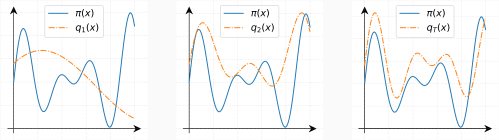
$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x$$

Can we sample from target distribution π ?

- YES, use naive Monte Carlo estimate (+ control variates)

$$\hat{\alpha}_n^{(\mathrm{mc})}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad X_1, \ldots, X_n \sim \pi$$

- NO, use **Adaptive Importance Sampling** with sampling policy $(q_i)_{i \geq 0}$



Evolution of sampling policy is AIS.

$$X_1 \sim q_0, \ldots, X_i \sim q_{i-1}$$

$$\hat{\alpha}_n^{(\mathrm{ais})}(f) = \frac{\sum_{i=1}^n w_i f(X_i)}{\sum_{i=1}^n w_i}$$

where the sequence $(w_i)_{i=1,\ldots,n}$ of **importance weights** is defined by

$$w_i = \pi(X_i)/q_{i-1}(X_i).$$

Adaptive Importance Sampling with Control Variates

AISCV estimate: Weighted Least Squares

Particles $X_i \sim q_{i-1}$ and weights $w_i = \pi(X_i)/q_{i-1}(X_i)$,

$$(\hat{\alpha}_n, \hat{\beta}_n) = \operatorname*{arg\,min}_{a \in \mathbb{R}, b \in \mathbb{R}^m} \sum_{i=1}^{n} w_i \left[f(X_i) - a - b^\top h(X_i) \right]^2 .$$

- (a) (Exact integration) whenever f is of the form $\alpha + \beta^\top h$ for some $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^m$, the **error is zero**, i.e., $\hat{\alpha}_n = \pi(f) = \int f \pi \, \mathrm{d}\lambda$.
- (b) (Quadrature Rule) $\hat{\alpha}_n = \sum_{i=1}^n v_{n,i} f(X_i)$, for **quadrature weights** $v_{n,i}$ **that do not depend on the function** f and that can be computed by a single weighted least squares procedure.
- (c) (Bayesian) it can be computed even when π **is known only up to a multiplicative constant**.
- (d) (post-hoc) CV can be brought into play in a ***post-hoc* scheme**, after generation of the particles and importance weights, and **this for any AIS algorithm**

Non-asymptotic error analysis

Residuals $\varepsilon = f - \alpha - \beta^\top h$ with $(\alpha, \beta) = \arg\min_{a,b} \int (f - a - b^\top h)^2 \pi \mathrm{d}\lambda$.

Assumptions

(A1) $\exists c \geq 1 : \forall x \in \mathbb{R}^d, \quad \pi(x) \leq c \cdot q_i(x)$.

(A2) $\sup_{x:\pi(x)>0} |h_j(x)| < \infty$ and $G = \mathbb{E}_\pi[hh^\top]$ invertible.

(A3) $\exists \tau > 0 : \forall t > 0, i \geq 1, \ \mathbb{P}[|w_i \varepsilon(X_i)| > t \mid \mathcal{F}_{i-1}] \leq 2 \exp(-t^2/(2\tau^2))$

Concentration inequality for AISCV estimate

Under assumptions, for any $\delta \in (0,1)$ and for all $n \geq C_1 c^2 B \log(10m/\delta)$, we have, with probability at least $1 - \delta$, that

$$\left| \hat{\alpha}_n^{(\mathrm{aiscv})}(f) - \pi(f) \right| \leq C_2 \sqrt{\log(10/\delta)} \frac{\tau}{\sqrt{n}} + C_3 c B \log(10m/\delta) \frac{\tau}{n},$$

where C_1, C_2, C_3 are some constants and $B = \sup_{x:\pi(x)>0} \|\hbar(x)\|_2^2$, $\hbar = G^{-1/2} h$.

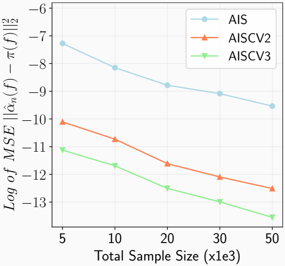
Synthetic examples: Gaussian Mixtures

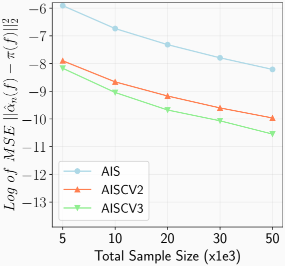
Similar framework as Cappé et al. (2008).

Integrand and Target: $f(x) = x, \pi_\Sigma(x) = 0.5\Phi_\Sigma(x-\mu) + 0.5\Phi_\Sigma(x+\mu)$ where $\mu = (1, \ldots, 1)^\top / 2\sqrt{d}, \Sigma = I_d/d$ and Φ_Σ is pdf $\mathcal{N}(0, \Sigma)$.

Sampling policy: Multivariate Student

Control variates: Stein method with $\varphi =$ polynomial with bounded degree





Gaussian mixture density: Logarithm of $\|\hat{\alpha}_n(f) - \pi(f)\|_2^2$ for $f(x) = x$ with target isotropic π_Σ with $d = 4$ (left), $d = 8$ (right).

Complexity rates for integration error

Definition: Root Mean Squared Error (RMSE)

The error δ_n of a procedure $\hat{\alpha}_n(f)$ that approximates $\pi(f)$ is

$$\delta_n = \mathbb{E}\left[|\hat{\alpha}_n(f) - \pi(f)|^2\right]^{1/2}$$

→ Lipschitz integrands[5], **optimal rate** in $O(n^{-1/2}n^{-1/d})$ (Novak, 2016)

OLS control variates (Portier and Segers, 2019)	$O(n^{-1/2}m^{-1/d})$
Determinantal sampling (Bardenet and Hardy, 2020)	$O(n^{-1/2}n^{-1/2d})$
Control Functionals (Oates et al., 2017)	$O(n^{-7/12})$
Cubic Stratification (Haber, 1966; Chopin and Gerber, 2022)	$O(n^{-1/2}n^{-1/d})$

[5] for integrand with s bounded derivatives, rate in $O(n^{-1/2}n^{-s/d})$

General view of Control Variates

Control Functionals

- Build surrogate function $\hat{f}$ with known integral $\pi(\hat{f})$
- Use centered variables $\hat{f}(X_i) - \pi(\hat{f})$ to derive the following enhanced

Monte Carlo estimate with control variates

$$\hat{\alpha}_n^{(CV)}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i) - \left(\hat{f}(X_i) - \pi(\hat{f})\right)\right\}$$

Approximation in $L_2(\pi)$

Let $(X_1, \ldots X_n) \sim \pi$. Suppose that $\hat{f}$ depends only on a surrogate sample $\tilde{X}_1, \ldots, \tilde{X}_N$ which is independent from $(X_1, \ldots X_n)$, then

$$\mathbb{E}\left[|\hat{\alpha}_n^{(CV)}(f) - \pi(f)|^2\right] \leq \frac{1}{n}\mathbb{E}\left[\int (f - \hat{f})^2 \mathrm{d}\pi\right].$$

Control Functionals examples

- **RKHS approximation:** (Oates, Girolami, and Chopin, 2017)

Ridge regression in Hilbert space $\mathcal{H}$

$$\hat{f} \in \arg\min_{\varphi \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} (f(\tilde{X}_i) - \varphi(\tilde{X}_i))^2 + \lambda \|\varphi\|_{\mathcal{H}}^2$$

- **Basis functions:** (Portier and Segers, 2019; Leluc et al., 2021)

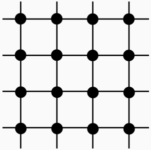
Use m basis functions $h_1, \ldots, h_m$ to fit OLS:

$$\hat{f} = \hat{\beta}_n^\top h, \qquad (\hat{\alpha}_n, \hat{\beta}_n) = \arg\min_{(\alpha,\beta) \in \mathbb{R} \times \mathbb{R}^m} \|f^{(n)} - \alpha \mathbb{1}_n - H\beta\|_2^2$$

- **Partitioning and Stratification:** (Chopin and Gerber, 2022)

$(\tilde{X}_1, \ldots, \tilde{X}_N)$ is the $(1/\ell)$-equidistant grid of $[0,1]^d$ with $N = \ell^d$, $\ell \geq 1$ and $(R_i)_{i=1,\ldots,N}$ is the partition of $[0,1]^d$ made of the rectangles.

$$\hat{f}(x) = \sum_{i=1}^{N} f(\tilde{X}_i) \mathbb{1}_{R_i}(x)$$



Nearest Neighbors

Control Neighbors

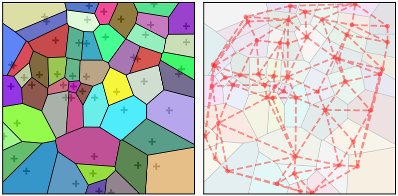
$$\hat{\alpha}_n^{(CVNN)}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\{ f(X_i) - \left(\hat{f}_n^{(i)}(X_i) - \pi(\hat{f}_n) \right) \right\}$$

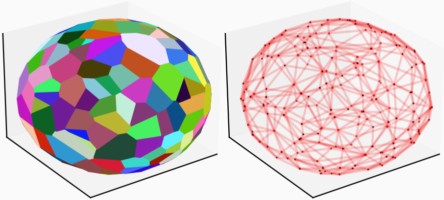
Leave-one-out Nearest Neighbors:

Take same sample $(X_1, \ldots, X_n)$ and define

$$\hat{f}_n(x) = \sum_{j=1}^{n} f(X_j)\mathbb{1}_{S_{n,j}}(x), \qquad \hat{f}_n^{(i)}(x) = \sum_{j \neq i} f(X_j)\mathbb{1}_{S_{n,j}^{(i)}}(x)$$

where $S_{n,j}$ are **Voronoï cells**





Control Neighbors properties

Control Neighbors

$$\hat{\alpha}_n^{(CVNN)}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i) - \left(\hat{f}_n^{(i)}(X_i) - \pi(\hat{f}_n)\right)\right\}$$

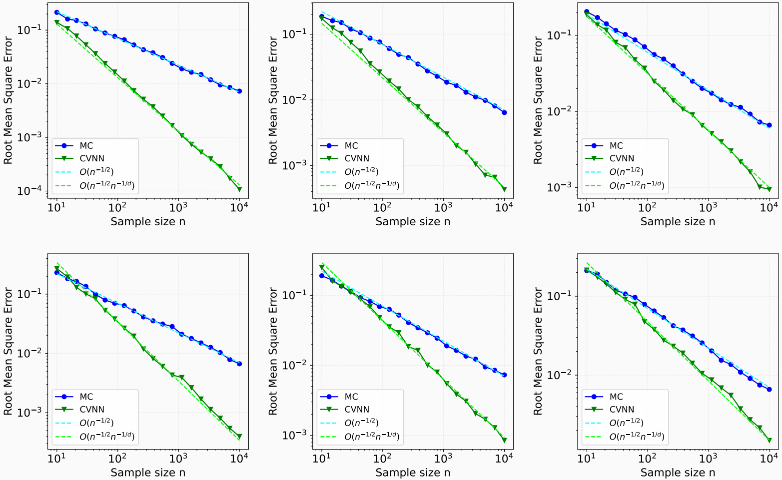
- (a) (Same framework as naive MC) does not require the existence of control variates with known integrals
- (b) (Quadrature Rule) $\hat{\alpha}_n = \sum_{i=1}^{n} w_{n,i} f(X_i)$, for **quadrature weights $w_{n,i}$ that do not depend on the function f.**
- (c) (Practical tool box) The weights $w_{n,i}$ are built using efficient nearest neighbors estimates (Bentley, 1975; Pedregosa et al., 2011)
- (d) (post-hoc) CVNN can be brought into play in a ***post-hoc* scheme** $\rightarrow$ include other sampling design like MCMC or AIS.

Complexity rate for integration error of Control Neighbors

$$\mathbb{E}\left[|\hat{\alpha}_n^{(CVNN)}(f) - \pi(f)|^2\right]^{1/2} \leq Cn^{-1/2}n^{-1/d}$$

Control Neighbors on synthetic integrands

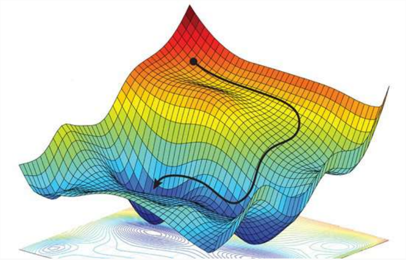
- $f_1(x_1, \dots, x_d) = \sin(\pi(\frac{2}{d} \sum_{i=1}^{d} x_i - 1))$ with $\pi = \mathbb{1}_{[0,1]^d}$
- $f_2(x_1, \dots, x_d) = \sin(\frac{\pi}{d} \sum_{i=1}^{d} x_i)$ with $\pi = \mathcal{N}_d(0, I_d)$



Error curves for f_1(top) and f_2(bottom) with $d \in \{2; 3; 4\}$

Part II: Optimize $\mathcal{F}$

Stochastic Optimization



1. **R. Leluc** and F. Portier. *Asymptotic Analysis of Conditioned Stochastic Gradient Descent. Transactions on Machine Learning Research, 2023* (Leluc and Portier, 2020)

2. **R. Leluc** and F. Portier. *SGD with Coordinate Sampling: Theory and Practice.* In *Journal of Machine Learning Research 23 (JMLR)*, (342):1–47, 2022. (Leluc and Portier, 2022)

Stochastic Optimization

Underlying optimization problem

Let $\mathcal{F} : \Theta \to \mathbb{R}$ be a general objective function.

- **Goal:**

$$\min_{\theta \in \Theta} \{\mathcal{F}(\theta) = \mathbb{E}_{z \sim \pi}[f(\theta, z)]\}$$

- **Constraints:** $\nabla \mathcal{F}$ is hard to compute (large-scale problems) or even intractable (black-box) !

Empirical Risk Minimization. $\hat{\mathcal{F}}(\theta) = n^{-1} \sum_{i=1}^{n} f_i(\theta)$ and true gradient, $n^{-1} \sum_{i=1}^{n} \nabla f_i(\theta)$ requires n evaluations, too heavy !

Stochastic Gradient Descent (Robbins and Monro, 1951)

$$\text{(SGD)}\ \theta_{t+1} = \theta_t - \gamma_{t+1} g_t \quad \text{with} \quad \mathbb{E}[g_t] = \nabla \mathcal{F}(\theta_t)$$

Ref: Robbins and Siegmund (1971); Bertsekas and Tsitsiklis (2000); Sacks (1958); Kushner and Clark (1978); Pelletier (1998); Benaïm (1999); Gadat et al. (2018); Moulines and Bach (2011); Bottou et al. (2018)

Limitations of SGD: choice of the learning rate (γ_t)

Conditioned-SGD

$$(\text{CSGD})\ \theta_{t+1} = \theta_t - \gamma_{t+1} C_t g_t$$

Research Questions (Part II)

- What condition on C_t for convergence? Asymptotic normality?
- How to leverage structure in data?

Existing methods (motivation)

- *2nd Order methods*: $C_t \approx \nabla^2 \mathcal{F}(\theta^\star)^{-1}$ or $C_t \approx \nabla^2 \mathcal{F}(\theta_t)^{-1}$

Stochastic Newton and Quasi-Newton (Byrd et al., 2016) and (L)BFGS methods (Liu and Nocedal, 1989; Moritz et al., 2016)

- *Fisher information matrix*: $C_t = F(\theta_t)$

Natural gradient (Amari, 1998; Kakade, 2002)

- *(Diagonal) Scalings*: $C_t = G_t^{-1/2}$; $G_{t+1} = G_t + g_t g_t^\top$

AdaGrad (Duchi et al., 2011), RMSProp (Tieleman et al., 2012), Adam (Kingma and Ba, 2014) and AMSGrad (Reddi et al., 2018)

From SGD...to Conditioned-SGD

Optimization problem

For general non-convex $\mathcal{F}$, find $\theta^\star \in \arg\min_{\theta\in\Theta} \{\mathcal{F}(\theta) = \mathbb{E}_\xi[f(\theta,\xi)]\}$

Central Limit Theorem CSGD

Under standard assumptions, if $C_t \to C$ almost surely then the iterates of CSGD satisfy

$$\frac{(\theta_t - \theta^\star)}{\sqrt{\gamma_t}} \rightsquigarrow \mathcal{N}(0, \Sigma_C), \quad \text{as } t \to +\infty.$$

- Optimal choice $C^\star = H^{-1}$ with $H = \nabla^2\mathcal{F}(\theta^\star)$ in the sense: $\Sigma_{C^\star} \preceq \Sigma_C$
- Practical procedure to achieve optimality $C_t \to C^\star$

SGD with Coordinate Sampling

(SCGD): Stochastic Coordinate Gradient Descent

$$(SCGD) \quad \theta_{t+1} = \theta_t - \gamma_{t+1} C(\zeta_{t+1}) g_{t+1}$$

with $C(k) = e_k e_k^T = Diag(0, \ldots, 0, 1, 0, \ldots, 0)$.
ζ_{t+1} is a random variable valued in $[\![1, d]\!]$.
$\rightarrow$ Reduction of computing cost
$\rightarrow$ 2 sources of randomness: noisy gradient g_t + random ζ_t

Research Questions and Contributions

• How to update the selecting policy ζ_{t+1} ?
$\rightarrow$ algorithm **MUSKETEER** to leverage the data structure and move along relevant directions.
• What condition on ζ_{t+1} for convergence ?
$\rightarrow$ analysis of the properties of $SCGD$ algorithms (convergence of the iterates, convergence of the policy, non-asymptotic bound)

Related work

- CD using $\mathcal{F}$ or true gradient $\nabla\mathcal{F}$ (Loshchilov et al., 2011; Richtárik and Takáč, 2016; Glasmachers and Dogan, 2013; Qu and Richtárik, 2016; Allen-Zhu et al., 2016; Namkoong et al., 2017)

- Most related idea: **Gauss-Southwell rule** to select the largest gradient coordinate to move the iterate (Nutini et al., 2015)

  $\rightarrow$ Here: stochastic g_t and ζ_t

- **Sparsification methods** (Alistarh et al., 2017; Wangni et al., 2018) , unbiased importance sampling estimate of the gradient

  $\rightarrow$ Here: no reweighting (biased) (conditioned gradient)

General framework and notation

- Only one coordinate ζ_{t+1} is selected: $\boxed{\theta_{t+1} = \theta_t - \gamma_{t+1} C(\zeta_{t+1}) g_{t+1}}$

$$\begin{cases} \theta_{t+1}^{(k)} = \theta_t^{(k)} & \text{if } k \neq \zeta_{t+1} \\ \theta_{t+1}^{(k)} = \theta_t^{(k)} - \gamma_{t+1} g_{t+1}^{(k)} & \text{if } k = \zeta_{t+1} \end{cases}$$

- The distribution of ζ_{t+1}, is the **coordinate sampling policy** and is given by the probability weights vector $p_t = (p_t^{(1)}, \dots, p_t^{(d)})$

$$\boxed{p_t^{(k)} = \mathbb{P}(\zeta_{t+1} = k | \mathcal{F}_t), \quad k \in [\![1, d]\!].}$$

- Not the same mean field as in usual SGD. Under conditional independence between g_{t+1} and ζ_{t+1}:

$$\mathbb{E}[C(\zeta_{t+1}) g_{t+1} | \mathcal{F}_t] = \mathrm{Diag}(p_t) \nabla \mathcal{F}(\theta_t)$$

Reinforcement Coordinate Sampling with MUSKETEER

MUltivariate
Stochastic
Knowledge
Extraction
Through
Exploration
Exploitation
Reinforcement

MUSKETEER

MUSKETEER may be seen as an **adaptive bandit** problem with

$$'arms = coordinates'$$

Alternate between 2 phases

- **Exploration phase (one for all)** (duration T)
  1. fix $p = p_t$, draw random coordinate $\zeta \sim p$ and noisy gradient g
  2. move iterate: $\theta^{(\zeta)} \leftarrow \theta^{(\zeta)} - \gamma g^{(\zeta)}$
  3. update gains of visited coordinates: $G^{(\zeta)} \leftarrow G^{(\zeta)} + g^{(\zeta)}/p^{(\zeta)}$

- **Exploitation phase (all for one)**
  1. share knowledge of the total gains
  2. update probability vector p_t with mixture

$$p_{t+1}^{(k)} = (1-\lambda)\frac{\exp(\eta|G_t^{(k)}|/t)}{\sum_{j=1}^{d}\exp(\eta|G_t^{(j)}|/t)} + \lambda\frac{1}{d}$$

Numerical Experiments: Zeroth-Order Optimization

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

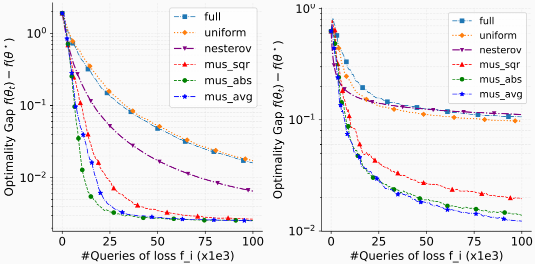
Special covariance structure

$X[:, k] \sim \mathcal{N}(0, \sigma_k^2 I_n)$ with $\sigma_k^2 = k^{-2}$ for $k \in [\![1, d]\!]$

- ZO gradient estimates:

(*finite differences*) $g_h(\theta, \xi) = \sum_{k=1}^{d} h^{-1}[f(\theta + he_k, \xi) - f(\theta, \xi)]e_k$

(*Nesterov*) $g_h(\theta, \xi) = h^{-1}[f(\theta + hU, \xi) - f(\theta, \xi)]U$ with $U \sim \mathcal{N}(0, I)$



Training Losses for Ridge regression and Logistic regression, obtained over 100 replications. Parameters $\gamma_t = 1/t$, $n = 10,000$, $d = 250$, $T = \lfloor \sqrt{d} \rfloor = 15$

Main results: MUSKETEER

Gradients might be biased

There exists constant $c \geq 0$ such that

$$\forall h > 0, \theta \in \mathbb{R}^p, \quad \| \mathbb{E}_\xi [g_h(\theta, \xi)] - \nabla \mathcal{F}(\theta) \| \leq ch.$$

$h \geq 0$ is a parameter controlling the bias with condition $h_t^2 = O(\gamma_t)$

Theoretical results

- The sequence of iterates $(\theta_t)_{t \geq 0}$ obtained by MUSKETEER satisfies $\nabla \mathcal{F}(\theta_t) \to 0$ almost surely as $t \to +\infty$.
- The MUSKETEER's coordinate policy $(p_t)_{t \in \mathbb{N}}$ converges weakly to the uniform distribution.
- Let $(\theta_t)_{t \in \mathbb{N}}$ obtained by MUSKETEER with $\gamma_t = \gamma / t$ then

$$\mathbb{E}\left[\mathcal{F}(\theta_t) - \mathcal{F}^\star\right] = O(1/t)$$

Conclusion

Conclusion & Perspectives

$$\textit{Integrate } \mathcal{F}(\theta) = \int_{\mathcal{X}} f(x)\pi_\theta(\mathrm{d}x) \to \textit{Optimize } \mathcal{F} \text{ with } \nabla\mathcal{F}$$

Takeaways.

- Non-asymptotic theory and practical procedures for Monte Carlo methods with control variates; Optimal convergence rates with nearest neighbors.
- Asymptotic analysis of Conditioned SGD methods; Theoretical and practical study of SGD with coordinate sampling.

Future work.

- Control variates for Markov chains; concentration inequality for CVNN
- Federated Learning applications of adaptive sampling.