Noisy stochastic gradients for price prediction

Rásonyi Miklós
Rényi Institute and ELTE, Budapest
Joint work with N. H. Chau, A. Lovas, É. Moulines,
S. Sabanis, K. Tikosi, Y. Zhang

Budapest, 26th November, 2021

Adaptive estimates

Stochastic gradient method

We wish to minimize $F(\theta) := E[f(\theta, X)]$ where $F : \mathbb{R}^d \to \mathbb{R}_+$, X random variable.
$H(\theta, x) := \partial_\theta f(\theta, x)$, let X_i be stationary with law equal to X, $i \in \mathbb{N}$.
Let us try the following alogorithm:

$$\hat{\theta}_{k+1} := \hat{\theta}_k - a_k H(\hat{\theta}_k, X_{k+1}).$$

Fixed gain: $a_k := \lambda$ or decreasing gain, e.g. $a_k = 1/k$ is typical.

A textbook example

We wish to minimize

$$E|\theta^T Z_n - Y_n|^2 + g(\theta)$$

in $\theta \in \mathbb{R}^d$ where $(Z_n, Y_n)_{n\in\mathbb{Z}} \in \mathbb{R}^{d+1}$ is a stationary process.
The function g is to enforce regularization.
This regression problem is omnipresent. The data sequence has no reason to be i.i.d. in general.
Markov property may hold but long memory may also kick in (econometric time series, telecommunication traffic).
Solution: stochastic gradient (Langevin) algorithm. Not necessarily convex functionals.

Kiefer-Wolfowitz variant

If f is not differentiable or the derivative is difficult to calculate then rather

$$\tilde{\theta}_{k+1} := \tilde{\theta}_k - a_k \frac{f(\tilde{\theta}_k + c_k, X_{k+1}) - f(\tilde{\theta}_k - c_k, X'_{k+1})}{2c_k}.$$

Use of random directions: SPSA (Spall, L. Gerencsér)
Typical: $a_k = 1/k$, $c_k = 1/\sqrt[6]{k}$.
Let θ^* be the (unique) minimizer.

Convergence and error estimate

Under weak conditions (stability, Lipschitz-continuity, mixing condition):

$$E|\hat{\theta}_k - \theta_*| \leq \frac{C}{\sqrt{k}}.$$

In the Kiefer-Wolfowitz case scantier literature:

$$E|\tilde{\theta}_k - \theta_*| \leq \frac{C}{\sqrt[3]{k}}.$$

When $a_k = \lambda$ fix:

$$E|\hat{\theta}_k - \theta_*| \leq C\sqrt{\lambda}.$$

Sampling based on the Langevin equation

Langevin algorithm

Langevin equation:

$$dL_t = -h(L_t)\,dt + dB_t,$$

where $h = \nabla F$; its stationary law:

$$\mu_* \sim e^{-F(u)}\,du.$$

Euler-approximation:

$$\bar{\theta}^{\lambda}_{k+1} = \bar{\theta}^{\lambda}_{k} - \lambda h(\bar{\theta}^{\lambda}_{k}) + \sqrt{\lambda/\beta}\,\xi_{k+1},$$

where ξ_k Gauss, i.i.d.
For small λ and large k this approximates μ_* well.

The solution of the Langevin equation tends to the stationary law at an exponential speed.
The error caused by λ is generically of the order $\sqrt{\lambda}$. Under stronger (convexity) assumption better estimates hold true.
Total variation norm is used:

$$||\mu - \nu||_{TV} = \sup_{|\phi|\leq 1} \left| \int_{\mathbb{R}^d} \phi(u)\mu(du) - \int_{\mathbb{R}^d} \phi(u)\nu(du) \right|,$$

$\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$.

Stochastic gradient Langevin algorithm

$$\theta_{k+1}^{\lambda} = \theta_k^{\lambda} - \lambda H(\theta_k^{\lambda}, X_{k+1}) + \sqrt{\lambda/\beta}\,\xi_{k+1},$$

where $h(\theta) = E[H(\theta, X_0)]$, $\beta > 0$: inverse temperature. We will let $\beta \to \infty$, $\lambda \to 0$, $k \to \infty$.

Then the method samples

$$\mu_* \sim e^{-\beta F(u)}\,du$$

which, for β large, means finding the minimum.

X_k: observed data or random sample from huge dataset.

Convergence analysis

Let $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, $\mathscr{C}(\mu, \nu)$ the set of all couplings.

$$\tilde{W}_p(\mu, \nu) := \left(\inf_{\pi \in \mathscr{C}(\mu,\nu)} \int_{\mathbb{R}^{2d}} |x - y|^p \pi(dx, dy) \right)^{1/p}, \; p \geq 1$$

$$\begin{aligned} W_1(\mu, \nu) &:= \inf_{\pi \in \mathscr{C}(\mu,\nu)} \int_{\mathbb{R}^{2d}} \max\{|x - y|, 1\} \pi(dx, dy) \\ &\leq \min\{C||\mu - \nu||_{TV}, \tilde{W}_1(\mu, \nu)\}. \end{aligned}$$

Known results

M. Raginsky, A. Rakhlin, M. Telgarsky: Non-convex learning via stochastic gradient Langevin dynamics: a non-asymptotic analysis, 2017.
Theorem.

$$\tilde{W}_2(\mathrm{Law}(\theta_k^\lambda), \mu_*) \leq \varepsilon$$

provided that

$$\lambda \leq c_1(\varepsilon/\ln(1/\varepsilon))^4, \; k \geq c_2 \frac{\ln^5(1/\varepsilon)}{\varepsilon^4}.$$

Upper estimate $\tilde{W}_2(\mathrm{Law}(\theta_k^\lambda), \mu_*)$ depends on k: $e^{-c\lambda k} + k\lambda^{5/4}$.

There are functions Δ, b such that

$$\langle H(\theta, x), \theta\rangle \geq \Delta(x)|\theta|^2 - b(x).$$

Expresses a certain degree of pulling effect towards the "centre".
Mixing conditions about X_t.

Theorem. If Δ, b are constants,

$$\tilde{W}_1(\mathrm{Law}(\theta_k^\lambda), \mu_*) \leq \varepsilon$$

provided that

$$\lambda \leq c_1 \varepsilon^2, \; k \geq c_2 \frac{\ln(1/\varepsilon)}{\varepsilon^2}.$$

Upper estimate independent of k: $e^{-c\lambda k} + \sqrt{\lambda}$.

Theorem. Law(θ_k^λ) converges to a limit μ_* in total variation as $k \to \infty$.
Two alternative sets of conditions:

1. $E[\Delta(X_0)] > 0$, b constant, H at most linear, X_0 bounded, satisfies large deviation-type estimates.
2. Δ constant, b, H polynomial in x (H at most linear in θ). Boundedness relaxed to a moment condition.

Markov chains in random environments

Polish spaces $\mathscr{X}, \mathscr{Y}$ with Borel sigma-fields $\mathfrak{B}, \mathfrak{A}$ Let $Q : \mathscr{Y} \times \mathscr{X} \times \mathfrak{B} \to [0,1]$ be a family of probabilistic kernels parametrized by $y \in \mathscr{Y}$, i.e. for all $A \in \mathfrak{B}$, $Q(\cdot,\cdot,A)$ is $\mathfrak{A} \otimes \mathfrak{B}$-measurable and for all $y \in \mathscr{Y}$, $x \in \mathscr{X}$, $A \to Q(y,x,A)$ is a probability on $\mathfrak{B}$.

Let X_t, $t \in \mathbb{N}$ be a $\mathscr{X}$-valued stochastic process such that

$$P(X_{t+1} \in A | \mathscr{F}_t) = Q(Y_t, X_t, A) \ P\text{-a.s.}, \ \ t \geq 0, \tag{1}$$

where the filtration is defined by

$$\mathscr{F}_t := \sigma(Y_j, \ j \in \mathbb{Z}; \ X_j, \ 0 \leq j \leq t), \ \ t \geq 0.$$

This is a Markov chain in a random environment.

Let us consider the problem of online nonlinear prediction of Z_n as a function of the p previous observations $Z_{n-1},\dots,Z_{n-p}$. $f_\theta : \mathbb{R}^{p\times m} \to \mathbb{R}^m$, $\theta \in \mathbb{R}^d$ is a parametric family of (non-linear) twice continuously differentiable functions, such as the output of a neural network. We seek to minimize the regularized mean-square error, that is,

$$U(\theta) = E[|Z_p - f_\theta(Z_{p-1},\dots,Z_0)|^2] + c|\theta|^2 \qquad (2)$$

for some $c > 0$. Under technical conditions, SGLD applies.
Online price prediction: 27 methods.
Lago, J., De Ridder, F. and De Schutter, B.: Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms", *Applied Energy*, 221:386–405, 2018.

Non-continuous case

In general $F(\theta) = E[f(\theta, X)]$ for some random variable X.
Often f is *not continuous*, hence $h(\theta) = \nabla F(\theta)$ does not admit an obvious random representation.
For instance: f can be an indicator function: minimizing the probability of an event.

An example from financial mathematics

An agent decides in which stock (s)he invests his/her money for the next trading period.
Changes in the prices: X_k, Y_k, $k \in \mathbb{N}$. The investor maximizes

$$EU(1_{X_{k-1}>\theta_1, Y_{k-1}\leq\theta_2} X_k + 1_{X_{k-1}\leq\theta_1, Y_{k-1}>\theta_2} Y_k)$$

in θ_1, θ_2.
U: functional expressing relation to risk.

Suggesting a new algorithm

Inspired by Kiefer-Wolfowitz algorithm:

$$\theta_{k+1}^{\lambda} = \theta_k^{\lambda} - \lambda \frac{f(\theta_k^{\lambda} + c_k, X_k) - f(\theta_k^{\lambda} - c_k, X_k')}{2c_k} + \sqrt{\lambda/\beta}\,\xi_{k+1}.$$

Let $a_k := 1/k$ and $c_k = k^{-\gamma}$ for some $\gamma > 0$.

$$\tilde{\theta}_{k+1} := \tilde{\theta}_k - a_k \frac{f(\tilde{\theta}_k + c_k, X_{k+1}) - f(\tilde{\theta}_k - c_k, X_{k+1}')}{2c_k}.$$

Under suitable (complicated) assumptions:

$$E|\tilde{\theta}_k - \theta_*| \le \frac{C}{k^{1/5}},$$

when $\gamma = 1/5$ is chosen.
Continuity sets must be polyhedral, Lipschitz-continuity in the average, smoothness assumptions, stability, dissipativity, global parameter set.

THANK YOU FOR YOUR ATTENTION!