Stochastic Gradient Descent in Continuous Time

Discrete and Continuous Data

Jonas Latz

School of Mathematical and Computer Sciences, Heriot-Watt University
Maxwell Institute for Mathematical Sciences
Edinburgh, UK

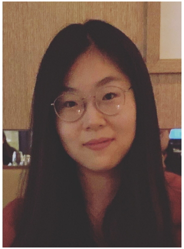
UQSay, Paris, September 29th 2022.

SGD in continuous time: discrete and continuous data

Related works: Jin, L., Liu, Schönlieb 2021: **A Continuous-time Stochastic Gradient Descent Method for Continuous Data**, under review.

L. 2021: **Analysis of stochastic gradient descent in continuous time**, Statistics and Computing 31, 39.

L. 2022: **Gradient flows and randomised thresholding: sparse inversion and classification**, under review.



Kexin Jin, Princeton

Chenguang Liu, Delft,

Carola-Bibiane Schönlieb, Cambridge

Funding: Engineering and Physical Sciences Research Council (EPSRC), Swindon, UK

Outline

Outline

Stochastic gradient descent - continuous time and discrete data

- Stochastic Gradient Descent with discrete data
- Continuous time models?
- Stochastic gradient process
- Longtime behaviour

Continuous data? - a motivation

Stochastic gradient descent - continuous time and continuous data

Illustrations

Conclusions

Optimisation problem: discrete data

- Consider an optimisation problem on $X := \mathbb{R}^K$; of the form

$$\theta^* \in \mathrm{argmin}_{\theta \in X} \bar{\Phi}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \Phi_i(\theta), \tag{OptP}$$

  where potentials $\bar{\Phi}, \Phi_i \in C^1(X; \mathbb{R}), i \in I := \{1, ..., N\}$ and (OptP) is well-defined.

Optimisation problem: discrete data

- Consider an optimisation problem on $X := \mathbb{R}^K$; of the form

$$\theta^* \in \mathrm{argmin}_{\theta \in X} \bar{\Phi}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \Phi_i(\theta), \tag{OptP}$$

  where potentials $\bar{\Phi}, \Phi_i \in C^1(X;\mathbb{R}), i \in I := \{1, ..., N\}$ and (OptP) is well-defined.

- Typical in statistical, imaging, and machine learning applications:
  - $\bar{\Phi}$: misfit between a model and a (big) data set
  - Φ_i: misfit between a model and the i-th partition of the data set

Gradient Descent and Stochastic Gradient Descent: discrete data

Gradient Descent (GD) for (OptP): [Cauchy; 1847]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla\bar{\Phi}(\theta_{k-1}), \qquad \nabla\bar{\Phi}(\theta_{k-1}) := \frac{1}{N}\sum_{i=1}^{N} \nabla\Phi_i(\theta_{k-1}).$$

Gradient Descent and Stochastic Gradient Descent: discrete data

Gradient Descent (GD) for (OptP): [Cauchy; 1847]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla\bar{\Phi}(\theta_{k-1}), \qquad \nabla\bar{\Phi}(\theta_{k-1}) := \frac{1}{N}\sum_{i=1}^{N} \nabla\Phi_i(\theta_{k-1}).$$

(convergence if $\bar{\Phi}$ is (strictly) convex and "step size" η_k is sufficiently small)

. . . .

Gradient Descent and Stochastic Gradient Descent: discrete data

Gradient Descent (GD) for (OptP): [Cauchy; 1847]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla \bar{\Phi}(\theta_{k-1}), \qquad \nabla \bar{\Phi}(\theta_{k-1}) := \frac{1}{N} \sum_{i=1}^{N} \nabla \Phi_i(\theta_{k-1}).$$

(convergence if $\bar{\Phi}$ is (strictly) convex and "step size" η_k is sufficiently small)

. . . .

Stochastic Gradient Descent (SGD) for (OptP): [Robbins & Monro; 1951]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla \Phi_{\boldsymbol{i}_k}(\theta_{k-1}), \qquad \underbrace{\boldsymbol{i}_k \sim \mathrm{Unif}(I)}_{(= \text{“subsampling”})} .$$

Gradient Descent and Stochastic Gradient Descent: discrete data

Gradient Descent (GD) for (OptP): [Cauchy; 1847]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla \bar{\Phi}(\theta_{k-1}), \qquad \nabla \bar{\Phi}(\theta_{k-1}) := \frac{1}{N} \sum_{i=1}^{N} \nabla \Phi_i(\theta_{k-1}).$$

(convergence if $\bar{\Phi}$ is (strictly) convex and "step size" η_k is sufficiently small)

Stochastic Gradient Descent (SGD) for (OptP): [Robbins & Monro; 1951]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla \Phi_{\boldsymbol{i}_k}(\theta_{k-1}), \qquad \underbrace{\boldsymbol{i}_k \sim \mathrm{Unif}(I)}_{(= \text{"subsampling"})} .$$

(convergence if $\Phi_1, \ldots, \Phi_N$ are strongly convex and "learning rate" $\eta_k \downarrow 0$ $(k \to \infty)$ slowly)

Stochastic Gradient Descent

- SGD constructs a Markov chain
- Stochastic properties hardly discussed [Benaïm; 1999][Dieuleveut et al.; 2017][Hu et al.; 2019]
  - Stationary measure, (Bayesian?) inference, and implicit regularisation
  - Ergodicity?
  - Speed of convergence?

    $\to$ this talk

Stochastic Gradient Descent

- SGD constructs a Markov chain
- Stochastic properties hardly discussed [Benaïm; 1999][Dieuleveut et al.; 2017][Hu et al.; 2019]
  - Stationary measure, (Bayesian?) inference, and implicit regularisation
  - Ergodicity?
  - Speed of convergence?
    $\rightarrow$ this talk
- Long-term goals
  - Construct more efficient stochastic optimisation algorithms
  - Understand random subsampling in SGD and other continuous-time methods; especially optimal convergence rates
  - Understand SGD in non-convex optimisation
  - Understand SGD with constant learning rates and implicit regularisation

Outline

In continuous time?

Idealisation and simplification of models through continuity assumption

- Usual modelling tool in many scientific disciplines (e.g., continuum mechanics,...)
- Recently also used in data science, machine learning, and algorithms
  - Ensemble Kalman Inversion [Schillings & Stuart; 2017, 2018][Blömker et al.; 2019]...
  - Continuum limits of graphs [Trillos & Sanz-Alonso; 2018] and in MCMC [Kuntz et al.; 2019]
  - PDE-based image reconstruction [Rudin et al.; 1992][Schönlieb; 2015]...
  - PDE-based data science **[Budd, van Gennip & L.; 2021]**[Kreusser & Wolfram; 2020]...
- continuous models tend to be easier to analyse: no numerical artefacts

A diffusion process?

Predominant model for SGD in continuous time: Diffusion process

- Idea: $\eta_k \approx 0 \Rightarrow$ gradient error is approximately Gaussian (CLT)
- Hence, $(\theta_k)_{k=1}^{\infty}$ can be represented by a diffusion process

$$\dot{\theta}(t) = -\nabla\bar{\Phi}(\theta(t)) + \Sigma(\theta(t))\dot{\mathrm{W}}_t \quad (t \geq 0), \qquad \theta(0) = \theta_0.$$

[Hu et al.; 2019][Li et al.; 2016, 2017, 2019][Mandt et al.; 2015, 2016, 2017][Wojtowytsch; 2021]

A diffusion process?

Predominant model for SGD in continuous time: Diffusion process

- Idea: $\eta_k \approx 0 \Rightarrow$ gradient error is approximately Gaussian (CLT)
- Hence, $(\theta_k)_{k=1}^{\infty}$ can be represented by a diffusion process

$$\dot{\theta}(t) = -\nabla \bar{\Phi}(\theta(t)) + \Sigma(\theta(t))\dot{\mathrm{W}}_t \quad (t \geq 0), \qquad \theta(0) = \theta_0.$$

[Hu et al.; 2019][Li et al.; 2016, 2017, 2019][Mandt et al.; 2015, 2016, 2017][Wojtowytsch; 2021]

Critique:

- for large η_k, the paths of $(\theta_k)_{k=1}^{\infty}$ are very different from a diffusion
  - preasymptotic phase and constant η_k not explained
- Diffusion does not actually explain subsampling in a continuous-time model
  - does not represent the discrete nature of the potential selection
  - needs access to $\bar{\Phi}$

Outline

Stochastic gradient descent - continuous time and discrete data

- Stochastic Gradient Descent with discrete data
- Continuous time models?
- Stochastic gradient process
- Longtime behaviour

Continuous data? - a motivation

Stochastic gradient descent - continuous time and continuous data

Illustrations

Conclusions

Observations and fundamental idea

- the update

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla \Phi_{\boldsymbol{i}_k}(\theta_{k-1}) \qquad \text{(discrete)}$$

  is a forward Euler discretisation of the gradient flow

$$\dot{\theta}(t) = -\nabla \Phi_{\boldsymbol{i}_k}(\theta(t)) \qquad \text{(continuous)}$$

- learning rate η_k has two different meanings
  (i) η_k is the step size of the gradient flow discretisation
  (ii) η_k determines the length of the time interval with which we switch the Φ_i

Observations and fundamental idea

- the update

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla \Phi_{\boldsymbol{i}_k}(\theta_{k-1}) \qquad \text{(discrete)}$$

  is a forward Euler discretisation of the gradient flow

$$\dot{\theta}(t) = -\nabla \Phi_{\boldsymbol{i}_k}(\theta(t)) \qquad \text{(continuous)}$$

- learning rate η_k has two different meanings
  (i) η_k is the step size of the gradient flow discretisation
  (ii) η_k determines the length of the time interval with which we switch the Φ_i

Idea.

Obtain a continuous time model for SGD, by

(i) let the step size go to 0, i.e. replace (discrete) by (continuous).

(ii) switch the potentials in the gradient flow at a rate of $1/\eta_k$

Switching of the potentials

control the switching of the potentials by a continuous-time Markov process (CTMP) $(\boldsymbol{i}(t))_{t\geq 0}$ on $I := \{1, ..., N\}$ ("index process")

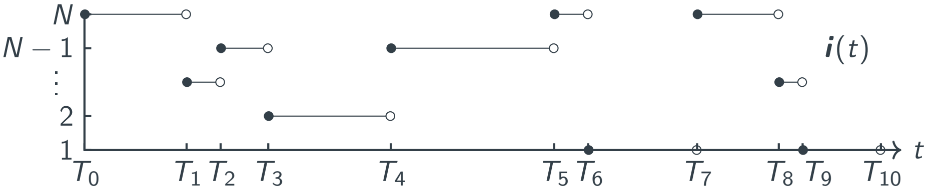


Figure: Cartoon of a CTMP

CTMPs 101

- $(\boldsymbol{i}(t))_{t\geq 0}$ is piecewise constant
- randomly jumps from one state to another after a random waiting time $\Delta \sim \pi_{\mathrm{wt}}(\cdot|t_0)$

Switching of potentials

Two versions: constant learning rate and decreasing learning rate

Switching of potentials

Two versions: constant learning rate and decreasing learning rate

(i) CTMP $(\boldsymbol{i}(t))_{t\geq 0}$ representing a constant learning rate $\eta_\bullet \equiv \eta > 0$

- constant learning rates are popular in practice
- $\pi_{\mathrm{wt}}(\cdot|t_0)$ is constant in time (indeed this will be an exponential distribution)

$(\boldsymbol{i}(t))_{t\geq 0}$ has constant transition rate matrix $A \in \mathbb{R}^{N\times N}$: $A_{i,j} := \begin{cases} \frac{1}{(N-1)\eta}, & \text{if } i \neq j, \\ -\frac{1}{\eta}, & \text{if } i = j. \end{cases}$

Switching of potentials

Two versions: constant learning rate and decreasing learning rate

(i) CTMP $(\boldsymbol{i}(t))_{t\geq 0}$ representing a constant learning rate $\eta_\bullet \equiv \eta > 0$

- constant learning rates are popular in practice
- $\pi_{\mathrm{wt}}(\cdot|t_0)$ is constant in time (indeed this will be an exponential distribution)

$(\boldsymbol{i}(t))_{t\geq 0}$ has constant transition rate matrix $A \in \mathbb{R}^{N\times N}$: $A_{i,j} := \begin{cases} \frac{1}{(N-1)\eta}, & \text{if } i \neq j, \\ -\frac{1}{\eta}, & \text{if } i = j. \end{cases}$

(ii) CTMP $(\boldsymbol{j}(t))_{t\geq 0}$ representing a decreasing learning rate $\eta_\bullet > 0,$ with $\eta_k \downarrow 0$ $(k \to \infty)$

- actually a chance of converging to the minimiser of $\bar{\Phi}$
- waiting times $\Delta \sim \pi_{\mathrm{wt}}(\cdot|t_0)$ get 'smaller' over time (in some sense)

$(\boldsymbol{j}(t))_{t\geq 0}$ has time-dependent transition rate matrix $B \in \mathbb{R}^{N\times N\times[0,\infty)}$: $B(t)_{i,j} := \begin{cases} \frac{1}{(N-1)H(t)}, & \text{if } i \neq j, \\ -\frac{1}{H(t)}, & \text{if } i = j, \end{cases}$

where $(H(t))_{t\geq 0}$ is continuously differentiable & interpolates $(\eta_k)_{k=1}^{\infty}$.

Stochastic gradient process

the Stochastic gradient process (SGP) is our continuous-time version of SGD

Definition. [L.; 2021]

We define the Stochastic gradient process...

(i) ...with constant learning rate (SGPC) by $(\theta(t))_{t\geq 0}$, which satisfies

$$\dot{\theta}(t) = -\nabla\Phi_{\boldsymbol{i}(t)}(\theta(t)) \quad (t \geq 0), \qquad \theta(0) = \theta_0.$$

(ii) ...with decreasing learning rate (SGPD) by $(\xi(t))_{t\geq 0}$, which satisfies

$$\dot{\xi}(t) = -\nabla\Phi_{\boldsymbol{j}(t)}(\xi(t)) \quad (t \geq 0), \qquad \xi(0) = \xi_0.$$

$(\theta(t))_{t\geq 0}$ and $(\xi(t))_{t\geq 0}$ are almost surely well-defined, if

Assumption [Lipschitz]. *For $i \in I : \Phi_i \in C^1(X, \mathbb{R})$ and $\nabla\Phi_i$ is Lipschitz continuous.*

Stochastic gradient process

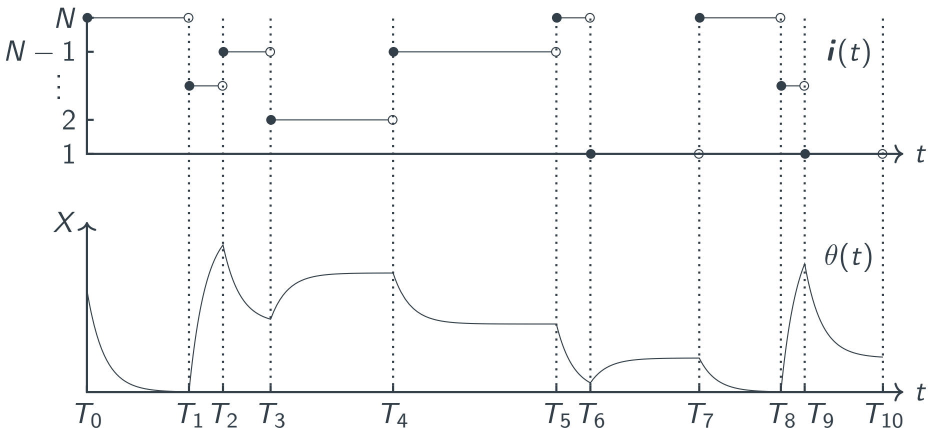


Figure: Cartoon of SGPC

Piecewise deterministic Markov processes

$(\theta(t), \boldsymbol{i}(t))_{t\geq 0}$, $(\xi(t), \boldsymbol{j}(t))_{t\geq 0}$ are piecewise deterministic Markov processes (PDMPs)

- 'a general class of non-diffusion stochastic models' [Davis; 1984, 1993]
- progression via deterministic dynamic (ODE) with jumps after random waiting times or when hitting a boundary
  [Bakhtin & Hurth; 2012][Benaïm et al.; 2012, 2015][Yin & Zhu; 2010]...
- used for stochastic modelling in engineering, computer science, and biology
  [Rudnicki & Tyran-Kamińska; 2017]
- used as a basis for non-reversible MCMC algorithms
  [Bierkens et al.; 2019][Fearnhead et al.; 2018][Power & Goldman; 2019],...

SGD vs. SGP

Gradient flow

Uniform sampling

Markov property

Learning rate

Approximation of deterministic gradient flow

SGD vs. SGP

Approximation of deterministic gradient flow

SGD with constant learning rate $\eta \approx 0$ approximates the 'exact' gradient flow

$$\frac{\mathrm{d}\zeta}{\mathrm{d}t} = -\nabla\bar{\Phi}(\zeta(t)), \qquad \zeta(0) = \theta_0.$$

Intuition:

- Euler scheme converges $\Rightarrow$ gradient flow
- law of large numbers (LLN):

$$\theta_k = \theta_0 - \big(\eta\nabla\Phi_{\boldsymbol{i}_1}(\theta_0) + \cdots + \eta\nabla\Phi_{\boldsymbol{i}_k}(\theta_{k-1})\big) \overset{(\eta\approx 0)}{\approx} \theta_0 - \underbrace{\big(\eta\nabla\Phi_{\boldsymbol{i}_1}(\theta_0) + \cdots + \eta\nabla\Phi_{\boldsymbol{i}_k}(\theta_0)\big)}_{\overset{\text{LLN}}{\approx}\eta k\bar{\Phi}(\theta_0)}$$

SGD vs. SGP

SGPC, with $\eta \approx 0$, also approximates the 'exact' gradient flow

Assumption `[Smooth]`. *For any $i \in I$, let $\Phi_i \in C^2(X;\mathbb{R})$ and let $\nabla\Phi_i, \mathrm{H}\Phi_i$ be continuous and bounded on bounded subsets of X.*

Theorem. **[L.; 2021]**

Let $\zeta(0) = \theta(0)$ and let Assumption `[Smooth]` hold, then $(\theta(t))_{t\geq 0} \to (\zeta(t))_{t \geq 0}$, weakly in $(C^0([0,\infty);X), \|\cdot\|_\infty)$, as $\eta \downarrow 0$.

Proof. Perturbed test function theory of [Kushner; 1984]. $\square$

SGD vs. SGP

Example. Let $\Phi_1(\theta) := (\theta - 1)^2/2$ and $\Phi_2(\theta) := (\theta + 1)^2/2$. $\Rightarrow \bar{\Phi}(\theta) = (\theta^2 + 1)/2$.

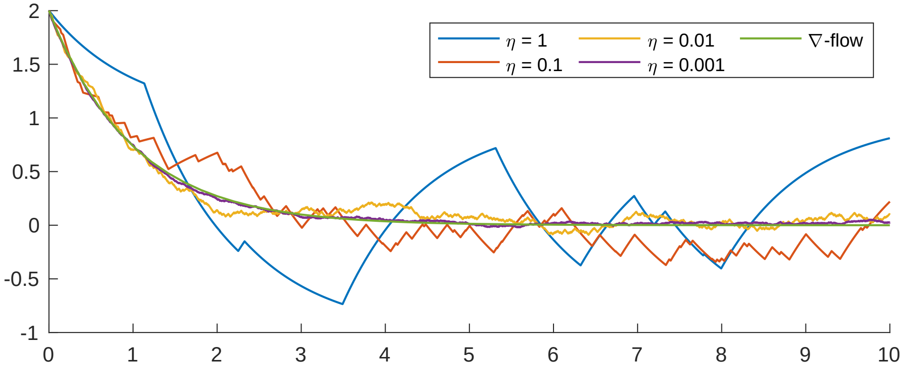


Figure: Exemplary realisations of SGPC and plot of precise gradient flow. Discretisation with `ode45`.

Outline

Stochastic gradient descent - continuous time and discrete data

- Stochastic Gradient Descent with discrete data
- Continuous time models?
- Stochastic gradient process
- Longtime behaviour

Continuous data? - a motivation

Stochastic gradient descent - continuous time and continuous data

Illustrations

Conclusions

Long-time behaviour of the Stochastic Gradient Process

Study long-time behaviour of the stochastic gradient processes, i.e., study

$$\mathbb{P}(\theta(t) \in \cdot), \qquad \mathbb{P}(\xi(t) \in \cdot) \qquad\qquad (t \gg 0 \text{ very large}).$$

- existence and uniqueness of stationary measures
- convergence to stationary measures and its speed
- SGPD: convergence to $\delta(\cdot - \theta^*)$, where $\theta^* \in \operatorname{argmin}_{\theta \in X} \bar{\Phi}(\theta)$

Preliminaries

Wasserstein distance

Let $q \in (0,1]$. Consider Wasserstein distance between $\pi, \pi' \in \mathrm{Prob}(X)$:

$$\mathrm{W}_q(\pi,\pi') := \inf_{H \in \mathrm{Coup}(\pi,\pi')} \int_{X \times X} \min\{1, \|\theta - \theta'\|_2^q\} H(\mathrm{d}\theta, \mathrm{d}\theta'),$$

$$\mathrm{Coup}(\pi,\pi') := \{G \in \mathrm{Prob}(X^2) : \quad G(\cdot \times X) = \pi, \quad G(X \times \cdot) = \pi'\}$$

- metrises weak convergence, i.e.

$$\mathrm{W}_q(\pi_n, \pi) \to 0, \text{ as } n \to \infty \qquad \Leftrightarrow \qquad \pi_n \to \pi, \text{ weakly, as } n \to \infty$$

Preliminaries

Assumption [Smooth]. *For any* $i \in I$, *let* $\Phi_i \in C^2(X; \mathbb{R})$ *and let* $\nabla\Phi_i, \mathrm{H}\Phi_i$ *be continuous and bounded on bounded subsets of* X.

Assumption [Convex]. *There is some* $\kappa > 0$, *with*

$$\left\langle \theta_0 - \theta_0', \nabla\Phi_i(\theta_0) - \nabla\Phi_i(\theta_0') \right\rangle \geq \kappa \|\theta_0 - \theta_0'\|^2 \qquad (\theta_0, \theta_0' \in X, i \in I),$$

i.e. Φ_i are strongly convex for $i \in I$.

Constant learning rate

Theorem. **[L.; 2021]**

Let Assumptions [Smooth] and [Convex] hold. Then, $(\theta(t), \boldsymbol{i}(t))_{t>0}$ has a unique stationary measure π_{C} on $(X \times I, \mathcal{B}X \otimes 2^I)$. Moreover, there exist $\kappa', c > 0$ and $q \in (0, 1]$, with

$$\mathrm{W}_q(\pi_{\mathrm{C}}(\cdot \times I), \mathbb{P}(\theta(t) \in \cdot | \theta_0, i_0)) \leq c \exp(-\kappa' t) \left(1 + \sum_{i \in I} \int_X \|\theta_0 - \theta'\|^q \pi_{\mathrm{C}}(\mathrm{d}\theta' \times \{i\})\right)$$

$$(i_0 \in I, \theta_0 \in X).$$

Constant learning rate

Theorem. [L.; 2021]

Let Assumptions [Smooth] and [Convex] hold. Then, $(\theta(t), \boldsymbol{i}(t))_{t>0}$ has a unique stationary measure π_{C} on $(X \times I, \mathcal{B}X \otimes 2^I)$. Moreover, there exist $\kappa', c > 0$ and $q \in (0,1]$, with

$$\mathrm{W}_q(\pi_{\mathrm{C}}(\cdot \times I), \mathbb{P}(\theta(t) \in \cdot | \theta_0, i_0)) \leq c \exp(-\kappa' t) \left(1 + \sum_{i \in I} \int_X \|\theta_0 - \theta'\|^q \pi_{\mathrm{C}}(\mathrm{d}\theta' \times \{i\})\right) \quad (i_0 \in I, \theta_0 \in X).$$

- convergence with exponential speed
- proof based on results by [Benaïm et al.; 2012][Cloez & Hairer; 2015]
- convexity assumption can be weakened (needs Hörmander Bracket condition)
- finding an analytical expression for π_{C} is probably hard / π_{C} might describe the implicit regularisation of SGPC

Illustrative example: stationary measures of SGPC

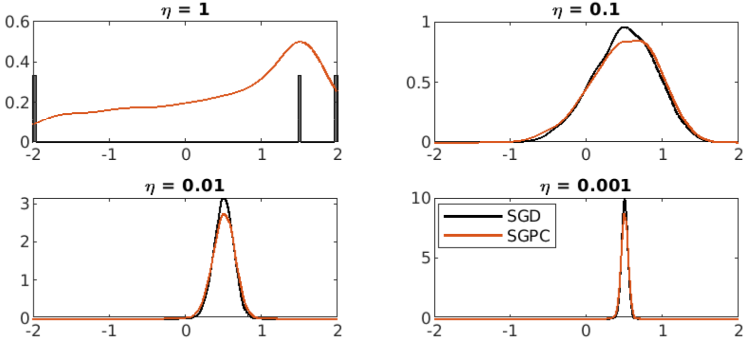


Figure: Kernel density estimates of $\mathbb{P}(\theta(10) \in \cdot|\theta(0) = -1.5) \approx \pi_{\mathrm{C}}$ (SGPC) and $\mathbb{P}(\theta_{10/\eta} \in \cdot|\theta_0 = -1.5)$ (SGD) based on $\eta \in \{1, 0.1, 0.01, 0.001\}$ using 10,000 samples each.

[**Example.** Let $N := 3$, i.e. $I := \{1, 2, 3\}$, and $X := \mathbb{R}$. We define the potentials $\Phi_1(\theta) := \frac{1}{2}(\theta + 2)^2$, $\Phi_2(\theta) := \frac{1}{2}(\theta - 1.5)^2$, $\Phi_3(\theta) := \frac{1}{2}(\theta - 2)^2$ $(\theta \in X)$. Here, $\mathrm{argmin}\bar{\Phi} = \{0.5\}$.]

Decreasing learning rate

Theorem. **[L.; 2021]**

Let Assumptions [Smooth] and [Convex] hold. Then, for any $\xi_0 \in X$ and $j_0 \in I$, we have

$$\mathrm{W}_1(\delta(\cdot - \theta^*), \mathbb{P}(\xi(t) \in \cdot | \xi_0, j_0)) \to 0 \qquad (t \to \infty).$$

Decreasing learning rate

Theorem. **[L.; 2021]**

Let Assumptions [Smooth] and [Convex] hold. Then, for any $\xi_0 \in X$ and $j_0 \in I$, we have

$$\mathrm{W}_1(\delta(\cdot - \theta^*), \mathbb{P}(\xi(t) \in \cdot | \xi_0, j_0)) \to 0 \qquad (t \to \infty).$$

- Convergence, but not really information about its speed
  - same problem exists for the diffusion model of SGD
- proof is significantly more involved
  - $(\xi(t), \boldsymbol{j}(t))_{t \geq 0}$ is inhomogeneous in time
  - rate matrix $B(\cdot)$ degenerates, as $t \to \infty$
  - uses results from [Benaïm et al.; 2012][Cloez & Hairer; 2015][Kushner; 1984]

Illustrative convergence plot of SGPD

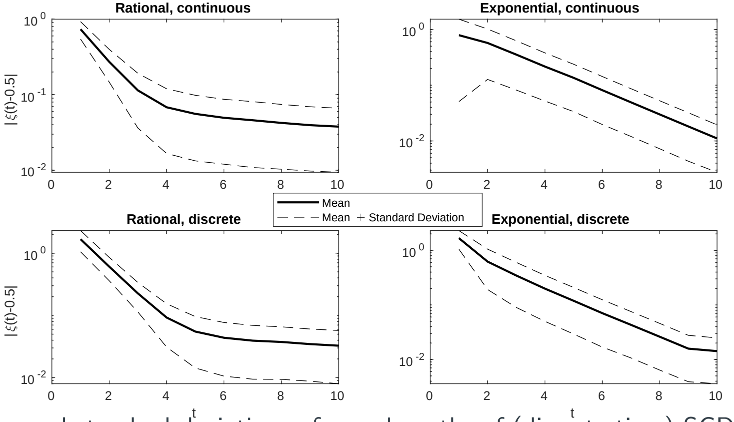


Figure: Mean error and standard deviations of sample paths of (discrete-time) SGD vs. (continuous-time) SGPD. Estimated using 10,000 samples. [Learning rates: $H(t) := (100t + 1)^{-1}$ (rational) and $H(t) := \exp(-t)$ (exponential)]

Outline

Optimisation problem: continuous data

Consider an optimisation problem on $X := \mathbb{R}^K$; of the form

$$\theta^* \in \mathrm{argmin}_{\theta \in X} \bar{\Phi}(\theta) := \int_S f(\theta, y)\pi(\mathrm{d}y), \tag{OptPCont}$$

with potentials $\bar{\Phi}, f(\cdot, y) \in C^1(X; \mathbb{R}), y \in S$, a compact space, and some general probability measure π on $(S, \mathcal{B}S)$.

Optimisation problem: continuous data

Consider an optimisation problem on $X := \mathbb{R}^K$; of the form

$$\theta^* \in \mathrm{argmin}_{\theta \in X} \bar{\Phi}(\theta) := \int_S f(\theta, y)\pi(\mathrm{d}y), \qquad \text{(OptPCont)}$$

with potentials $\bar{\Phi}, f(\cdot, y) \in C^1(X; \mathbb{R}), y \in S$, a compact space, and some general probability measure π on $(S, \mathcal{B}S)$.

Multiple applications

- robust optimisation: control of uncertain systems
- functional data analysis/machine learning: physics-informed neural networks, adaptive imaging
- variational inference: optimise **E**vidence **L**ower **BO**und
- spatial model for a high-dimensional discrete problem: image reconstruction with large data availability

Physics-informed Neural Networks

Example.

Let $\mathcal{L} : H \to H'$ be a differential operator on appropriate spaces H, H' of functions from $S \to \mathbb{R}$ and $g \in H'$. Moreover, let H'' represent functions: $\partial S \to \mathbb{R}$ and let $B : H \to H''$ be another operator. PDE:

$$\text{Find } u \in H : \qquad \begin{cases} \mathcal{L}u(x) = g(x) & (x \in S^\circ) \\ Bu(x) = 0 & (x \in \partial S). \end{cases}$$

Physics-informed Neural Networks

Example.

Let $\mathcal{L} : H \to H'$ be a differential operator on appropriate spaces H, H' of functions from $S \to \mathbb{R}$ and $g \in H'$. Moreover, let H'' represent functions: $\partial S \to \mathbb{R}$ and let $B : H \to H''$ be another operator. PDE:

$$\text{Find } u \in H : \qquad \begin{cases} \mathcal{L}u(x) = g(x) & (x \in S^\circ) \\ Bu(x) = 0 & (x \in \partial S). \end{cases}$$

Physics-informed Neural Networks:

- let $U : X \to H$ be an appropriate function (deep neural network with weights and biases in X)
- solve: $\min_{\theta \in X} \int_S \left(\mathcal{L}U(\theta)(x) - g(x)\right)^2 \mathrm{d}x + \int_{\partial S} \left(BU(\theta)(x)\right)^2 \mathrm{d}x$

Physics-informed Neural Networks

Example.

Let $\mathcal{L} : H \to H'$ be a differential operator on appropriate spaces H, H' of functions from $S \to \mathbb{R}$ and $g \in H'$. Moreover, let H'' represent functions: $\partial S \to \mathbb{R}$ and let $B : H \to H''$ be another operator. PDE:

$$\text{Find } u \in H : \qquad \begin{cases} \mathcal{L}u(x) = g(x) & (x \in S^\circ) \\ Bu(x) = 0 & (x \in \partial S). \end{cases}$$

Physics-informed Neural Networks:

- let $U : X \to H$ be an appropriate function (deep neural network with weights and biases in X)
- solve: $\min_{\theta \in X} \int_S (\mathcal{L}U(\theta)(x) - g(x))^2 \,\mathrm{d}x + \int_{\partial S} (BU(\theta)(x))^2 \,\mathrm{d}x$
  (Here: $\pi := \mathrm{Unif}(S) \otimes \mathrm{Unif}(\partial S)$. Usually: replace integral by a quadrature rule)

Stochastic Gradient Descent: continuous data

How do we solve (OptPCont)?

$$\theta^* \in \mathrm{argmin}_{\theta \in X} \bar{\Phi}(\theta) := \int_S f(\theta, y)\pi(\mathrm{d}y) \qquad \text{(OptPCont)}$$

Stochastic Gradient Descent: continuous data

How do we solve (OptPCont)?

$$\theta^* \in \mathrm{argmin}_{\theta \in X} \bar{\Phi}(\theta) := \int_S f(\theta, y)\pi(\mathrm{d}y) \qquad \text{(OptPCont)}$$

Stochastic Gradient Descent (SGD) for (OptPCont): [Robbins & Monro; 1951]

for $k = 1, 2, \ldots$:

$$\theta_k \leftarrow \theta_{k-1} - \eta_k \nabla f(\theta_{k-1}, y_k), \qquad y_k \sim \pi.$$

- no need to compute the integral
- epochs are infinite

Outline

Stochastic gradient descent - continuous time and discrete data

Continuous data? - a motivation

Stochastic gradient descent - continuous time and continuous data

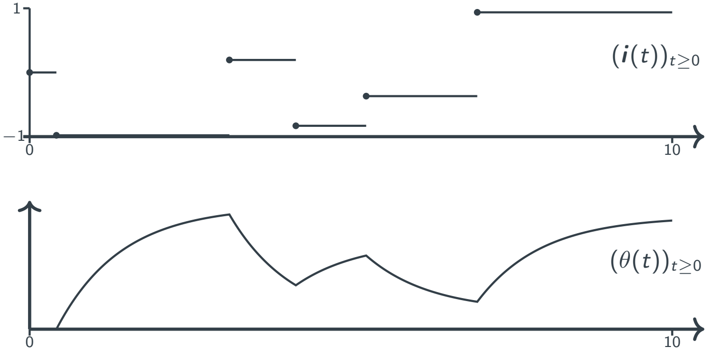
- Idea
- Index processes and the stochastic gradient process with continuous data
- Longtime behaviour

Illustrations

Conclusions

Stochastic gradient process with continuous data

Easy, right? Define the Stochastic Gradient Process as in the discrete data case with $(\boldsymbol{i}(t))_{t\geq 0}$ being now a pure Markov jump process on, say, $S := [-1, 1]$ with stationary measure π.



Stochastic gradient process with continuous data

Easy, right? Define the Stochastic Gradient Process as in the discrete data case with $(\boldsymbol{i}(t))_{t\geq 0}$ being now a pure Markov jump process on, say, $S := [-1, 1]$ with stationary measure π.

Actually,

- $(\boldsymbol{i}(t))_{t\geq 0}$ ignores spatial information in S
  - $(\boldsymbol{i}(t))_{t\geq 0}$ essentially samples independently from π
  - Complex sampling patterns?
- Implicit regularisation?
- The measure π could be complicated and independent samples not be available
  - obtain samples from MCMC in Bayesian inference or statistical physics simulations

Stochastic gradient process with continuous data

Easy, right? Define the Stochastic Gradient Process as in the discrete data case with $(\boldsymbol{i}(t))_{t\geq 0}$ being now a pure Markov jump process on, say, $S := [-1, 1]$ with stationary measure π.

Actually,

- $(\boldsymbol{i}(t))_{t\geq 0}$ ignores spatial information in S
  - $(\boldsymbol{i}(t))_{t\geq 0}$ essentially samples independently from π
  - Complex sampling patterns?
- Implicit regularisation?
- The measure π could be complicated and independent samples not be available
  - obtain samples from MCMC in Bayesian inference or statistical physics simulations

Idea: Allow for more general index processes

Allow for more general index processes

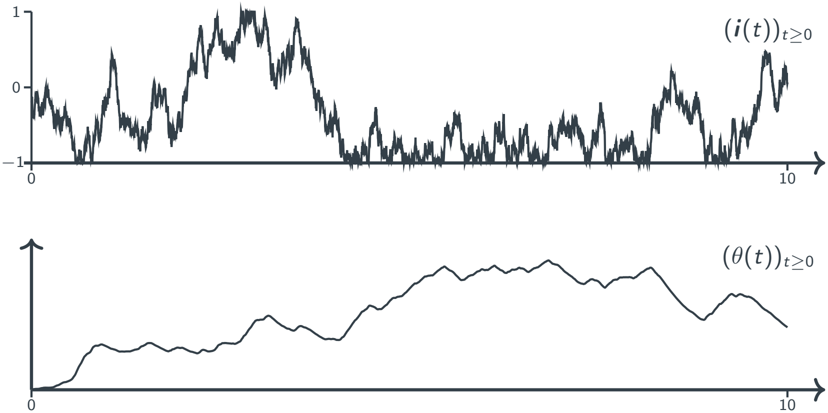


Figure: Stochastic gradient process with reflected diffusion index process

Outline

Index process

Definition and assumption [Index].

[Jin, L., Liu, Schönlieb, 2021]

Let $(V_t)_{t\geq 0}$ be a Feller process on $(\Omega, \mathcal{F}, (\mathcal{F}_t))_{t\geq 0}, (\mathbb{P}_x)_{x\in S})$. We assume the following:

(i) $(V_t)_{t\geq 0}$ admits a unique invariant measure π.

(ii) For any $x \in S$, there exist a family $(V_t^x)_{t\geq 0}$ and a stationary version $(V_t^\pi)_{t\geq 0}$ defined on the same probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that, $(V_t^x)_{t\geq 0} = (V_t)_{t\geq 0}$ in $\mathbb{P}_x$ and $(V_t^\pi)_{t\geq 0} = (V_t)_{t\geq 0}$ in $\mathbb{P}_\pi$.

(iii) Let $T^x := \inf\{t \geq 0 \mid V_t^x = V_t^\pi\}$ be a stopping time. There exist constants $C, \delta > 0$ such that for any $t \geq 0$, $\sup_{x\in S} \tilde{\mathbb{P}}(T^x \geq t) \leq C\exp(-\delta t)$.

We refer to $(V_t)_{t\geq 0}$ as index process.

$\Rightarrow$ $(V_t)_{t\geq 0}$ is exponentially ergodic: $d_{\mathrm{TV}}(\pi, \mathbb{P}_x(V_t \in \cdot)) \leq C\exp(-\delta t)$, $x \in S, t \geq 0$.

Examples of index processes

Example: *Markov pure jump process*

$(\boldsymbol{i}(t))_{t\geq 0} =: (V_t)_{t\geq 0}$ on $S \subseteq \mathbb{N}$ as given in the first part of this talk

- also $S = \mathbb{N}$ or $S \subsetneq \mathbb{R}$ being a compact interval are possible

Example: *Reflected Lévy processes*

$(V_t)_{t\geq 0}$ being a reflected Lévy process on a compact interval $S \subsetneq \mathbb{R}$

- e.g., a reflected Brownian motion

Also, finite products of such reflected Lévy processes on compact intervals

Stochastic gradient process with constant learning rate

Definition. [Jin, L., Liu, Schönlieb; 2021]

Let $(V_t)_{t\geq 0}$ be an index process and let $\varepsilon > 0$. Then, $(\theta_t^\varepsilon)_{t\geq 0}$ given by

$$\frac{\mathrm{d}\theta_t^\varepsilon}{\mathrm{d}t} = -\nabla f(\theta_t^\varepsilon, V_{t/\varepsilon}), \qquad \theta_0^\varepsilon = \theta_0 \in X,$$

is called stochastic gradient process with constant learning rate.

$(V_t, \theta_t^\varepsilon)_{t\geq 0}$ is well-defined and Markovian under Assumptions `[Index]`, `[Smooth2]`.

Stochastic gradient process with constant learning rate

Definition. [Jin, L., Liu, Schönlieb; 2021]

Let $(V_t)_{t\geq 0}$ be an index process and let $\varepsilon > 0$. Then, $(\theta_t^\varepsilon)_{t\geq 0}$ given by

$$\frac{\mathrm{d}\theta_t^\varepsilon}{\mathrm{d}t} = -\nabla f(\theta_t^\varepsilon, V_{t/\varepsilon}), \qquad \theta_0^\varepsilon = \theta_0 \in X,$$

is called stochastic gradient process with constant learning rate.

$(V_t, \theta_t^\varepsilon)_{t\geq 0}$ is well-defined and Markovian under Assumptions `[Index]`, `[Smooth2]`.

Assumption `[Smooth2]`. Let $f(x, y) \in \mathcal{C}^2(X \times S, \mathbb{R})$.

1. $\nabla_x f$, $H_x f$ are continuous and bounded on $X' \times S$ where $X' \subset X$ is bounded.
2. $\nabla_x f(x, y)$ is Lipschitz in x and the Lipschitz constant is uniform for $y \in S$.
3. For $x \in X$, $f(x, \cdot)$ and $\nabla_x f$ are integrable w.r.t to the probability measure $\pi(\cdot)$.

Learning rate? ε?

- $\varepsilon > 0$ is a scaling parameter that we use to control the 'learning rate'

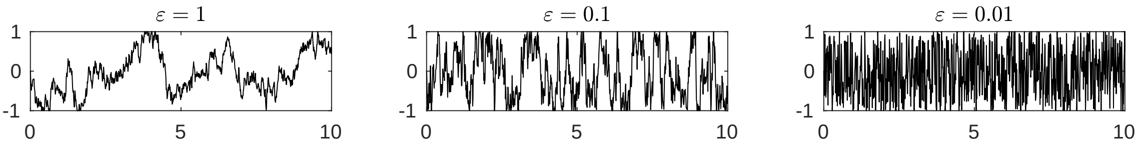


Figure: $(V_{t/\varepsilon})_{t\geq 0}$, where $(V_t)_{t\geq 0}$ is a reflected Brownian motion.

Idea: Small $\varepsilon \Rightarrow$ short correlation length in $(V_t)_{t\geq 0} \Rightarrow$ small learning rate

Learning rate? ε?

- $\varepsilon > 0$ is a scaling parameter that we use to control the 'learning rate'

- Approximation of the full gradient flow $(\zeta_t)_{t \geq 0}$, where

$$\frac{\mathrm{d}\zeta_t}{\mathrm{d}t} = -\nabla \int_S f(\zeta_t, y)\pi(\mathrm{d}y), \qquad \zeta_0 = \theta_0$$

Theorem. [Jin, L., Liu, Schönlieb; 2021]

Let Assumptions [Index], [Smooth2] hold. Then,

$$\int_0^\infty \exp(-t)\min\{1, \sup_{0 \leq s \leq t} \|\theta_t^\varepsilon - \zeta_t\|\}\mathrm{d}t \to 0, \text{ weakly, as } \varepsilon \downarrow 0.$$

Proof. Similar ideas to the approximation result with discrete data; harder as $(V_{t/\varepsilon})_{t \geq 0}$ is not necessarily tight with respect to $\varepsilon > 0$. Uses results from [Kushner; 1984; 1990].

Stochastic gradient process with decreasing learning rate

Idea: Let $\varepsilon \downarrow 0$ slowly over time.

Stochastic gradient process with decreasing learning rate

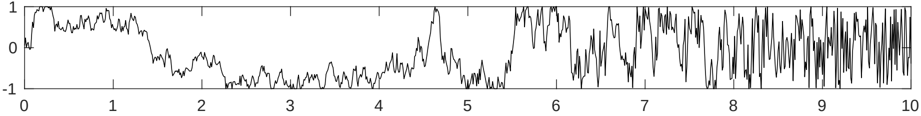
Idea: Let $\varepsilon \downarrow 0$ slowly over time.

Definition. **[Jin, L., Liu, Schönlieb; 2021]**

Let $\beta(s) := \int_0^s \mu(t)\mathrm{d}t$ with $\mu : [0,\infty) \to (0,\infty)$ non-decreasing, continuously differentiable with $\lim_{t\to\infty} \mu(t) = \infty$ very slowly. Moreover, let $(V_t)_{t\geq 0}$ be a suitable index process. Then, we define the stochastic gradient process with decreasing learning rate by $(\xi_t)_{t\geq 0}$ through

$$\frac{\mathrm{d}\xi_t}{\mathrm{d}t} = -\nabla f(\xi_t, V_{\beta(t)}), \qquad \xi_0 = \theta_0 \in X.$$

Well-defined, if `[Index]` and `[Smooth2]` are satisfied.



Outline

Longtime behaviour

Summary

[Jin, L., Liu, Schönlieb; 2021]

Results are fairly similar to the discrete data case:

Assumption Convex2: Require $x \mapsto f(x, y)$ be strongly convex, uniformly in $y \in S$

SGPC: Existence of a unique stationary measure of $(V_{t/\varepsilon}, \theta_t^\varepsilon)_{t \geq 0}$. Obtain exponential ergodicity in Wasserstein-1 distance

SGPD: Obtain convergence to the Dirac measure concentrated in $\theta^* \in \mathrm{argmin}_{\theta \in X} \int f(\theta, y)\pi(\mathrm{d}y)$ in Wasserstein-1 distance

Techniques: Lyapunov theory, weak Harris theorem [Cloez & Hairer; 2015]

Outline

Example: *Polynomial regression with functional data*

Data: Let $S := [-1, 1]$. We observe a function $g : S \to \mathbb{R}$, which is given by

$$g(y) = \underbrace{\sin(\pi y)}_{=:\Theta(y)} + \underbrace{\Xi(y)}_{\text{Gaussian noise}} \qquad (y \in S)$$

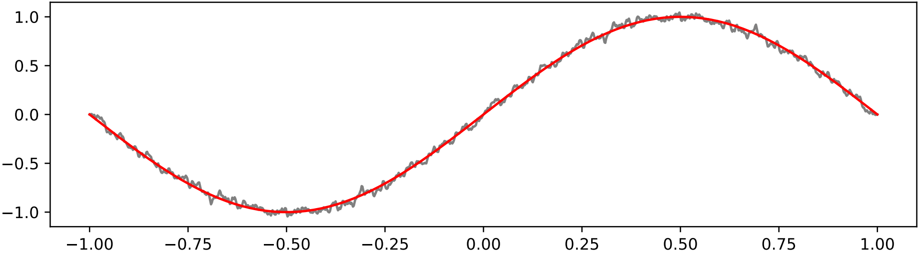


Figure: True function Θ (red) and noisy observation g (grey) in the polynomial regression example.

Example: *Polynomial regression with functional data*

Data: Let $S := [-1, 1]$. We observe a function $g : S \to \mathbb{R}$, which is given by

$$g(y) = \underbrace{\sin(\pi y)}_{=:\Theta(y)} + \underbrace{\Xi(y)}_{\text{Gaussian noise}} \qquad (y \in S)$$

Task

Reconstruct $\Theta : S \to \mathbb{R}$ on a polynomial basis $(\ell_k)_{k=1}^K$. In particular, minimise

$$\bar{\Phi}(\theta) := \frac{1}{2} \int_{[-1,1]} \left(g(y) - \sum_{k=1}^{K} \theta_k \ell_k(y) \right)^2 \mathrm{d}y + \frac{\alpha}{2} \|\theta\|_2^2 \qquad (\theta \in X),$$

Subsampled potential $f(\theta, y) := \frac{1}{2} \left(g(y) - \sum_{k=1}^{K} \theta_k \ell_k(y) \right)^2 + \frac{\alpha}{2} \|\theta\|_2^2 \qquad (\theta \in X, y \in S).$

Algorithmic setting

General

- Note that f satisfies the convexity assumption
- Study SGPC to learn about convergence and implicit regularisation

Algorithmic setting

General

- Note that f satisfies the convexity assumption
- Study SGPC to learn about convergence and implicit regularisation

Time-stepping of coupled dynamical system

- Considered dynamics: Reflected diffusion, Markov pure jump process with independently sampled jumps, and discrete SGD
- Discretise gradient flows with implicit midpoint rule with step size $= 0.1$
- Discretise index processes: Euler-Maruyama discretisation of diffusion with trivial reflection at boundary, precise sampling from Markov pure jump process with step size $= 0.01$

Error trajectory

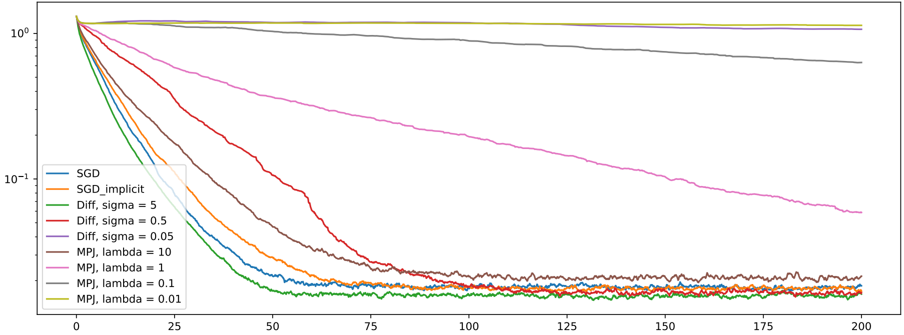


Figure: Relative error trajectory between the estimated polynomial and true function Θ; compare the function at 1000 points in S. Plot shows the mean over 100 error estimates. λ is the parameter of the exponential waiting time distribution. σ is the standard deviation of the Brownian motion before reflection.

Reconstruction errors

Method	Parameters	Mean of $\mathrm{rel\_err}_{N,(\cdot)}$	$\pm$ StD
SGD	$\eta_{(\cdot)} = 0.1$	$1.844 \cdot 10^{-2}$	$\pm 4.012 \cdot 10^{-3}$
SGD implicit	$\eta_{(\cdot)} = 0.1$	$1.719 \cdot 10^{-2}$	$\pm 3.939 \cdot 10^{-3}$
SGPC with reflected diffusion index process	$\sigma = 5$	$1.586 \cdot 10^{-2}$	$\pm 4.038 \cdot 10^{-3}$
	$\sigma = 0.5$	$1.587 \cdot 10^{-2}$	$\pm 2.979 \cdot 10^{-3}$
	$\sigma = 0.05$	$4.637 \cdot 10^{-2}$	$\pm 8.776 \cdot 10^{-2}$
SGPC with Markov pure jump index process	$\lambda = 10$	$2.100 \cdot 10^{-2}$	$\pm 6.049 \cdot 10^{-3}$
	$\lambda = 1$	$3.427 \cdot 10^{-2}$	$\pm 1.105 \cdot 10^{-2}$
	$\lambda = 0.1$	$3.866 \cdot 10^{-2}$	$\pm 1.142 \cdot 10^{-2}$
	$\lambda = 0.01$	$3.178 \cdot 10^{-1}$	$\pm 2.124 \cdot 10^{-1}$

Table: Mean and standard deviation of the relative error of the methods at the final point of their trajectory. In particular, sample mean and sample standard deviation of $j \mapsto \mathrm{rel\_err}_{N,j}$, with $N = 5 \cdot 10^4$, computed over 100 independent runs.

Discussion

- Ignoring the very slowly moving processes, all processes quickly reached an equilibrium state
- Interestingly, the SGPC with reflected diffusion appears to beat the other methods
  - implicit variance reduction due to large discrepancy between samples in S?
  - implicit regularisation of reflected diffusion especially effective?
- Computational cost of all methods in this example is fairly equivalent

Outline

Take-home messages

- we introduced SGP – a continuous-time model for SGD with discrete and continuous subsampling
- captures most properties of SGD
  - gradient flow structure, uniform subsampling, Markov property, learning rates/switching rate, approximates deterministic gradient flows
- The subsampling can be ‘essentially independent’ or following a Feller process
  - Allows for more general data sources and complex sampling patterns
- SGPC converges to a unique stationary measure π_{C} at exponential speed
- SGPD converges to $\delta(\cdot - \theta^*)$

Where do we go from here?

- Can we reach exponential convergence in SGPD?
- Develop efficient practical algorithms from SGP
- Mildly non-convex/non-smooth optimisation $\Rightarrow$ Recent preprint: **[L. 2022]**
  - Sparse (ℓ_1-)regularisation via randomised splitting
  - Classification via randomised Allen–Cahn equation
- SGD in 'very' non-convex optimisation
  - learning rate acts similar to a temperature in simulated annealing
- introduce subsampling in other continuous-time algorithms
- understand statistical properties of π_{C}
  - seems related to a posterior density [Mandt et al.; 2017]

HERIOT WATT UNIVERSITY

Jonas Latz web: www.latzplacian.org twitter: @latzplacian

Outline

Stochastic gradient descent - continuous time and discrete data

Continuous data? - a motivation

Stochastic gradient descent - continuous time and continuous data

Illustrations

Conclusions

Future research direction

SGP in practice

(i) discretise gradient flows $\dot{\theta}(t) = -\nabla\Phi_i(\theta(t))$, $\theta(0) = \theta_0$ for several $i \in I, \theta_0 \in X$

How do we discretise the gradient flows to retain the same ergodic behaviour?

(ii) discretise CTMPs $(\boldsymbol{i}(t))_{t\geq 0}$, $(\boldsymbol{j}(t))_{t\geq 0}$, $(V_t)_{t\geq 0}$

SGP in practice

(i) discretise gradient flows $\dot{\theta}(t) = -\nabla\Phi_i(\theta(t))$, $\theta(0) = \theta_0$ for several $i \in I, \theta_0 \in X$

(ii) discretise CTMPs $(\boldsymbol{i}(t))_{t\geq 0}$, $(\boldsymbol{j}(t))_{t\geq 0}$, $(V_t)_{t\geq 0}$

- Exact sampling of $(\boldsymbol{i}(t))_{t\geq 0}$, $(\boldsymbol{j}(t))_{t\geq 0}$ using algorithm by [Gillespie; 1977] : needs to sample waiting times from $\pi_{\mathrm{wt}}(\cdot|t_0)$
  - sampling from exponential distribution in case of $(\boldsymbol{i}(t))_{t\geq 0}$
  - more complicated in case of $(\boldsymbol{j}(t))_{t\geq 0}$

SGP in practice

(i) discretise gradient flows $\dot{\theta}(t) = -\nabla\Phi_i(\theta(t))$, $\theta(0) = \theta_0$ for several $i \in I, \theta_0 \in X$

(ii) discretise CTMPs $(\boldsymbol{i}(t))_{t\geq 0}$, $(\boldsymbol{j}(t))_{t\geq 0}$, $(V_t)_{t\geq 0}$

- Exact sampling of $(\boldsymbol{i}(t))_{t\geq 0}$, $(\boldsymbol{j}(t))_{t\geq 0}$ using algorithm by [Gillespie; 1977] : needs to sample waiting times from $\pi_{\mathrm{wt}}(\cdot|t_0)$
  - sampling from exponential distribution in case of $(\boldsymbol{i}(t))_{t\geq 0}$
  - more complicated in case of $(\boldsymbol{j}(t))_{t\geq 0}$
- The SGD-way: use fixed waiting times and sample from $\mathrm{Unif}(I)$
  - representation is quite imprecise, but might do the job
  - continuous time modelling step backwards
- How accurate do we need to discretise a, say, reflected diffusion?

SGD, Stochastic Proximal Point, SVRG, SAG, SAGA,...?

Retrieving well-known algorithms from SGP

- choose deterministic waiting times in the discretisation of the CTMP
- choose particular time stepping schemes for the gradient flows
  - forward Euler $\Rightarrow$ SGD [Robbins & Monro; 1951]
  - backward Euler $\Rightarrow$ Stochastic Proximal Point [Bertsekas; 2011]
  - forward Euler + control variate (or a multistep method?) $\Rightarrow$ SVRG [Johnson & Zhang; 2013], SAG [Schmidt et al.; 2017], SAGA [Defazio et al.; 2014]
  - higher order scheme $\Rightarrow$ higher order SGD-type method [Song et al.; 2018]
- Can we do better?