Derivatives of Stochastic Gradient Descent in parametric optimization

Franck Iutzeler, Edouard Pauwels, Samuel Vaiter

NeurIPS 2024

Analyzing the backpropagation through SGD is the same as analyzing *pertubated* SGD

$F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is smooth and **strongly convex** in x, P probability measure

Wait? But Why? → stochastic bilevel optimization, (model-agnostic) meta-learning, hyperparameter selection, etc.

Parametric stochastic optimization problem (Opt)

$$x^\star(\theta) = \arg\min_{x\in\mathbb{R}^d} F(x,\theta) := \mathbb{E}_{\xi\sim\mathrm{P}}[f(x,\theta;\xi)]$$

x: weights θ: parameter ξ: randomness

$x_k(\theta) \to_{k\to+\infty} x^\star(\theta)$ (under various assumptions)

Stochastic Gradient Descent (SGD)

$$x_0(\theta) \in \mathbb{R}^d$$

initialization (map)

$$x_{k+1}(\theta) = x_k(\theta) - \eta_k \nabla_x f(x_k(\theta),\theta;\xi_{k+1})$$

stepsize

i.i.d. rv following P

chain rule

Derivatives (Jacobian) of Stochastic Gradient Descent (SGD')

$$\partial_\theta x_{k+1}(\theta) = \partial_\theta x_k(\theta) - \eta_k \nabla^2_{xx} f(x_k(\theta),\theta;\xi_{k+1})\partial_\theta x_k(\theta) - \eta_k \nabla^2_{x\theta} f(x_k(\theta),\theta;\xi_{k+1})$$

Jacobian of $\theta \mapsto x_k(\theta)$

Hessian of $x \mapsto f(x,\theta,\xi)$

Cross-derivatives of $(x,\theta) \mapsto f(x,\theta,\xi)$

implicit function theorem

Jacobian of the solution: $\partial_\theta x^\star(\theta)$ solution of the linear system in D

$$\nabla^2_{xx}F(x^\star(\theta),\theta)D + \nabla^2_{x\theta}F(x^\star(\theta),\theta) = \mathbb{E}_{\xi\sim\mathrm{P}}\left[\nabla^2_{xx}f(x^\star(\theta),\theta;\xi)D + \nabla^2_{x\theta}f(x^\star(\theta),\theta;\xi)\right] = 0$$

Our issue

Do we have $\lim_{k\to+\infty} \partial_\theta x_k(\theta) = \partial_\theta x^\star(\theta)$?

Intuition from the quadratic case $f(x,\theta;\xi) = \frac{1}{2}\langle x, Q(\xi)x\rangle + \langle x, B(\xi)\theta\rangle$

$$\partial_\theta x_{k+1}(\theta) = \partial_\theta x_k(\theta) - \eta_k\left(Q(\xi_{k+1})\partial_\theta x_k(\theta) + B(\xi_{k+1})\right)$$

Stochastic reformulation (Opt') [Arbel-Mairal '22]

$$\partial_\theta x^\star(\theta) = \arg\min_{D\in\mathbb{R}^{d\times m}} \mathbb{E}_{\xi\sim\mathrm{P}}\left[\left\langle \frac{1}{2}\nabla^2_{xx}f(x^\star(\theta),\theta;\xi)D + \nabla^2_{x\theta}f(x^\star(\theta),\theta;\xi), D\right\rangle\right]$$

Our key insight

(SGD') is **inexact** (SGD) on (Opt')

General rewrite of (SGD')

$$\partial_\theta x_{k+1}(\theta) = \partial_\theta x_k(\theta) - \eta_k \nabla^2_{xx} f(x^\star(\theta),\theta;\xi_{k+1})\partial_\theta x_k(\theta) - \eta_k \nabla^2_{x\theta} f(x^\star(\theta),\theta;\xi_{k+1}) + e_{k+1}$$

"error term"

$$e_{k+1} = \eta_k\left(\nabla^2_{xx}f(x^\star(\theta),\theta;\xi_{k+1}) - \nabla^2_{xx}f(x_k(\theta),\theta;\xi_{k+1})\right)\partial_\theta x_k(\theta) + \eta_k\left(\nabla^2_{x\theta}f(x^\star(\theta),\theta;\xi_{k+1}) - \nabla^2_{x\theta}f(x_k(\theta),\theta;\xi_{k+1})\right)$$

Auxiliary result: convergence of inexact (SGD)

Stochastic optimization problem

$$w^\star = \arg\min_{w\in\mathbb{R}^d} F(w) := \mathbb{E}_{\xi\sim\mathrm{P}}[f(w;\xi)]$$

Inexact (SGD)

$$w_{k+1} = w_k - \eta_k(\nabla_w g(w_k;\xi_{k+1}) + e_{k+1})$$

General assumptions (Assumption G)

a) *Differentiability*: $g(\,\cdot\,;\xi)$ is C^1 with L-Lipschitz gradient ($\forall\xi$)

b) *Strong convexity*: $\exists w^\star, \mu > 0, \forall w, \langle w - w^\star, \mathbb{E}[\nabla_w g(w;\xi)\rangle \geq \mu\|w - w^\star\|^2$

c) *Variance control*: $\exists\sigma \in \mathbb{R}_+, \mathbb{E}[\|\nabla_w g(x;\xi)\|^2] \leq \sigma^2$

Inexact assumptions (Assumption I)

$\forall k, \mathbb{E}[\|e_{k+1}\|^2] \leq B_k^2$ for some (B_k) deterministic nonincreasing sequence (relative to the natural filtration associated to ξ_k)

Proposition (Convergence of inexact (SGD))

Under Assumptions G and I, set (η_k) a positive, nonincreasing, nonsummable sequence with $\eta_0 \leq \frac{\mu}{4L^2}$, and $D_k = \sqrt{\mathbb{E}[\|w_k - w^\star\|^2]}$ Then, for all $k \geq 0$,

$$D_{k+1}^2 \leq (1-\mu\eta_k)D_k^2 + 2\eta_k^2(B_k^2 + 2\sigma^2) + 2\eta_k B_k D_k$$

Stepsize	Errors	Noise	Result
$\eta_k \to \eta \geq 0$	$B_k \to B \propto \sqrt{\eta}$	$\sigma^2 \geq 0$	$\limsup D_k \propto \sqrt{\eta}$
$\eta_k = \frac{2\mu}{\mu^2 k + 8L^2}$	$B_k = 0$	$\sigma^2 \geq 0$	$D_k^2 = \mathcal{O}\left(\frac{\log(k+8\kappa^2)}{k+8\kappa^2}\right)$
$\eta_k = \frac{2\mu}{\mu^2 k + 8L^2}$	$B_k^2 = \mathcal{O}\left(\frac{\log(k+8\kappa^2)}{k+8\kappa^2}\right)$	$\sigma^2 \geq 0$	$D_k^2 = \mathcal{O}\left(\frac{\log(k+8\kappa^2)^2}{k+8\kappa^2}\right)$
$\eta_k = \eta < \frac{1}{2\mu}$	$B_k^2 = \mathcal{O}\left((1-\frac{\mu\eta}{2})^k\right)$	$\sigma^2 = 0$	$D_k^2 = \mathcal{O}\left(k(1-\frac{\mu\eta}{2})^k\right)$

Main result: convergence of (SGD')

Objective assumptions (Assumption O)

a) *Differentiability*: $f(\,\cdot\,,\cdot\,;\xi)$ is C^2 with M-Lipschitz Hessian ($\forall\xi$)

b) *Smoothness*: $\nabla_x f(\,\cdot\,,\theta,\xi)$ is L-Lipschitz and $\nabla_x f(x,\cdot\,,\xi)$ is L'-Lipschitz ($\forall x,\theta,\xi$)

c) *Strong convexity*: $f(\,\cdot\,,\theta,\xi)$ is μ-strongly convex ($\forall\theta,\xi$)

Noise assumptions (Assumption N) [Bottou et al., 2018; Gower et al., 2019]

a) *Variance control*: there exists $\sigma \geq 0$ such that for all $\theta \in \mathbb{R}^m$

$$\mathbb{E}[\|\nabla_x f(x^\star(\theta),\theta;\xi)\|^2] \leq \sigma^2$$ standard var. control

$$\mathbb{E}[\|\nabla^2_{xx}f(x^\star(\theta),\theta;\xi)\partial_\theta x^\star(\theta) + \nabla^2_{x\theta}f(x^\star(\theta),\theta;\xi)\|^2] \leq \sigma^2$$ **derivatives** var. control

b) *Integrability*: $f(x,\theta;\,\cdot\,)$ and $\nabla_x f(x,\theta;\,\cdot\,)$ are P-integrable

Theorem (Convergence of (SGD') towards (Opt'))

Under Assumptions O and N, set $\kappa = L/\mu$ the condition number, and (η_k) a positive, nonincreasing, nonsummable sequence with $\eta_0 \leq \frac{\mu}{4L^2}$. Then:

- General estimates: *setting $\eta = \lim_{k\to+\infty} \eta_k$, we have*

$$\limsup_{k\to+\infty}\mathbb{E}[\|\partial_\theta x_k(\theta) - \partial_\theta x^\star(\theta)\|^2] \leq \frac{4\sigma^2\eta}{\mu}\left(1 + \frac{3M(1+M_0)}{\mu}\right)^2$$

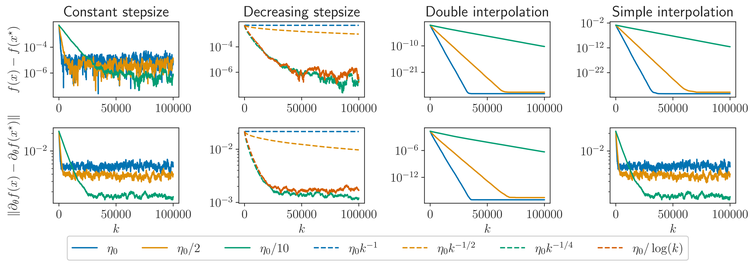
- Sublinear rate: *If $\eta_k = \frac{1}{\mu}\frac{2}{k+8\kappa^2}$, then*

$$\mathbb{E}[\|\partial_\theta x_k(\theta) - \partial_\theta x^\star(\theta)\|^2] = \mathcal{O}\left(\frac{\log(k+8\kappa^2)^2}{k+8\kappa^2}\right)$$

- Interpolation regime: *If $\sigma = 0$ and $\eta_k = \eta_0$ for all k, then*

$$\mathbb{E}[\|\partial_\theta x_k(\theta) - \partial_\theta x^\star(\theta)\|^2] = \mathcal{O}\left(k\left(1 - \frac{1}{8\kappa^2}\right)^k\right)$$

Numerical experiments



Ordinary Least square

$$x^\star(\theta) = \arg\min_{x\in\mathbb{R}^d} F(x,\theta) := \frac{1}{2m}\sum_{\xi=1}^m (a_\xi^\top x - b(\theta)_\xi)^2$$

General estimates & decreasing stepsize: $\theta \in \mathbb{R}^m, b(\theta) = \theta$

Simple interpolation setting: $\theta = A\zeta, \zeta \in \mathbb{R}^d$ and $b(\theta) = \theta$ (no autodiff through $\theta = A\zeta$)

Double interpolation setting: $\theta \in \mathbb{R}^d$ and $b(\theta) = A\theta$

References

[Arbel-Mairal, '22] Amortized Implicit Differentiation for Stochastic Bilevel Optimization. *ICML*.

[Bolte et al., '22] Automatic differentiation of nonsmooth iterative algorithms. *NeurIPS*.

[Bottou et al., '18] Optimization methods for large-scale machine learning. *SIAM Rev*.

[Chen et al., '87] Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stoch. Process. their Appl.*

[Gilbert, '92] Automatic differentiation and iterative processes. *Optim. Methods Softw.*

[Gower et al., '19] SGD: General analysis and improved rates. *ICML*.

[Iutzeler et al., '24] Derivatives of Stochastic Gradient Descent in parametric optimization. *NeurIPS*.