Controlling the Flow: Stability and Convergence for Stochastic Gradient Descent with Decaying Regularization

Sebastian Kassing, Simon Weißmann, Leif Döring

The Thirty-Ninth Annual Conference on Neural Information Processing Systems

Smooth & Convex Optimization:

Task: Minimize a differentiable function $f : \mathcal{X} \to \mathbb{R}$, where $(\mathcal{X}, \langle \cdot, \cdot \rangle_{\mathcal{X}})$ is a separable real Hilbert space.

Assumptions:

- f is convex,
- f is L-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$ and
- $\mathrm{argmin}_{x \in \mathcal{X}} f(x) \neq \emptyset$.

Minimum-norm solution: We denote by $x_* \in \mathrm{argmin}_{x \in \mathcal{X}} f(x)$ the minimum-norm solution, i.e. a minimum with $\|x_*\|_{\mathcal{X}} \leq \|\hat{x}\|_{\mathcal{X}}$ for all $\hat{x} \in \mathrm{argmin}_{x \in \mathcal{X}} f(x)$.

Regularized Stochastic Gradient Descent

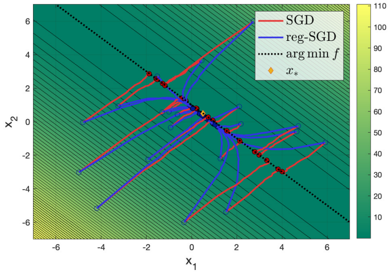
At iteration we have access to an unbiased estimator

$$\widehat{\nabla f(X_{k-1})} = \nabla f(X_{k-1}) + D_k$$

Assumptions:

- Unbiased: $\mathbb{E}[D_k \mid \mathcal{F}_{k-1}] = 0$
- ABC-condition:

$$\mathbb{E}\left[\|D_k\|^2 \mid \mathcal{F}_{k-1}\right] \leq A(f(X_{k-1}) - f(x_*)) + B\|\nabla f(X_{k-1})\|^2 + C$$



Stochastic Gradient Descent:

$$X_k = X_{k-1} - \gamma_k\big(\nabla f(X_{k-1}) + D_k\big),$$

Regularized Stochastic Gradient Descent:

$$X_k = X_{k-1} - \gamma_k\big(\nabla f(X_{k-1}) + \lambda_k X_{k-1} + D_k\big),$$

where $\gamma_k, \lambda_k \to 0$.

General last iterate almost sure convergence

Strategy: Balance the error $\|X_n - x_*\| \leq \|X_n - x_{\lambda_n}\| + \|x_{\lambda_n} - x_*\|$, where x_λ denotes the unique minimum of the strongly-convex function $f_\lambda(x) := f(x) + \frac{\lambda}{2}\|x\|^2$.

Theorem 2.1

Let $(\gamma_n)_{n\in\mathbb{N}_0}$ and $(\lambda_n)_{n\in\mathbb{N}_0}$ adapted (possibly random) sequences that are uniformly bounded from above. Assume that almost surely $\lambda_n \to 0$ (decreasingly) and that

$$\sum_{n\in\mathbb{N}} \gamma_n \lambda_n = \infty, \quad \sum_{n\in\mathbb{N}} \gamma_n^2 < \infty, \quad \text{and} \quad \sum_{n\in\mathbb{N}} \gamma_n \lambda_n \big(\|x_*\|_{\mathcal{X}}^2 - \|x_{\lambda_n}\|_{\mathcal{X}}^2\big) < \infty.$$

Then $\lim_{n\to\infty} \|X_n - x_*\| = 0$ almost surely.

General last iterate L^2-convergence

Strategy: Balance the error $\|X_n - x_*\| \leq \|X_n - x_{\lambda_n}\| + \|x_{\lambda_n} - x_*\|$, where x_λ denotes the unique minimum of the strongly-convex function $f_\lambda(x) := f(x) + \frac{\lambda}{2}\|x\|^2$.

Theorem 2.2

Let $(\gamma_n)_{n\in\mathbb{N}_0}$ and $(\lambda_n)_{n\in\mathbb{N}_0}$ be deterministic sequences of positive reals. Assume that $\lambda_n \to 0$ (decreasingly) and that

$$\sum_{n\in\mathbb{N}} \gamma_n\lambda_n = \infty, \quad \gamma_n = o(\lambda_n), \quad \text{and} \quad \lambda_n - \lambda_{n-1} = o(\gamma_n\lambda_n).$$

Then $\lim_{n\to\infty} \mathbb{E}[\|X_n - x_*\|^2] = 0$.

L^2-convergence rates

Choose

$$\gamma_n = C_\gamma (n+1)^{-q} \quad \text{and} \quad \lambda_n = C_\lambda (n+1)^{-p}.$$

Theorem 2.3

Assume that $p \in (0, \frac{1}{2}]$ and $q \in (p, 1-p]$ and, if $q = 1-p$, additionally assume that $2 C_\lambda C_\gamma > 1-q$. Then it holds that $\lim_{k\to\infty} \mathbb{E}[\|X_n - x_*\|^2] = 0$ and

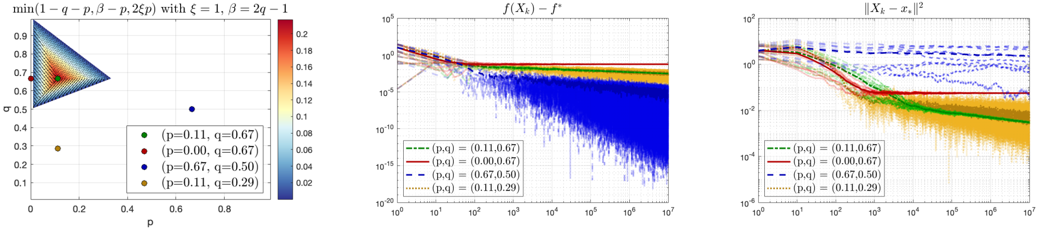
(i) $\mathbb{E}[f(X_n) - f(x_*)] \in \mathcal{O}(n^{-\min(p, q-p)})$,

(ii) $\mathbb{E}[\|X_n - x_{\lambda_n}\|^2] \in \mathcal{O}(n^{-\min(1-q-p, q-2p)})$ for $p \in (0, \frac{1}{3})$ and $q \in (2p, 1-p)$.

Optimizing the rates yield: $\mathbb{E}[f(X_n) - f(x_*)] \in \mathcal{O}(n^{-\frac{1}{3}})$ for $p = \frac{1}{3}$ and $q = \frac{2}{3}$.

Toy example

The PL-inequality implies $\|x_\lambda - x_*\| \lesssim \lambda^{1/4}$, see Maulen-Soto, Fadili, Attouch (2024).



- $f(X) = \frac{1}{2}(X_1 + X_2 - 1)^2$
- $D_k \sim \mathcal{N}(0, 0.01)$