Last iterate convergence of SGD for Least-Squares in the Interpolation regime

Aditya Varre*, Loucas Pillaud-Vivien*, Nicolas Flammarion*

NeurIPS, October 19, 2021.

* TML - EPFL

Problem Setting

- **Least-Square:** A stream of i.i.d samples $(x_i, y_i)_{i=1}^T$ from an unknown distribution ρ. We want to minimize the population risk:

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_\rho \left(\langle \theta, x \rangle_{\mathcal{H}} - y\right)^2 ,$$

where $\theta, x \in \mathcal{H}$, (possibly infinite dimensional) Hilbert space and $y \in \mathbb{R}$.

Problem Setting

- **Least-Square:** A stream of i.i.d samples $(x_i, y_i)_{i=1}^{T}$ from an unknown distribution ρ. We want to minimize the population risk:

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_{\rho}\left(\langle \theta, x\rangle_{\mathcal{H}} - y\right)^2,$$

where $\theta, x \in \mathcal{H}$, (possibly infinite dimensional) Hilbert space and $y \in \mathbb{R}$.

- **We study the SGD algorithm:**

$$\theta_{t+1} = \theta_t - \gamma\left(\langle \theta_t, x_t\rangle_{\mathcal{H}} - y_t\right)x_t$$

Problem Setting

- **Least-Square:** A stream of i.i.d samples $(x_i, y_i)_{i=1}^T$ from an unknown distribution ρ. We want to minimize the population risk:

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_\rho \left(\langle \theta, x\rangle_{\mathcal{H}} - y\right)^2 ,$$

where $\theta, x \in \mathcal{H}$, (possibly infinite dimensional) Hilbert space and $y \in \mathbb{R}$.

- **We study the SGD algorithm:**

$$\theta_{t+1} = \theta_t - \gamma \left(\langle \theta_t, x_t\rangle_{\mathcal{H}} - y_t\right) x_t$$

- **Aim: bound the excess risk.** Denote $\theta_* := \operatorname{argmin}_{\theta\in\mathcal{H}} \mathcal{R}(\theta)$, we bound the excess risk of the estimator given by the T-th iterate:

$$\mathbb{E}\mathcal{R}(\theta_T) - \mathcal{R}(\theta_*)$$

Last Iterate of SGD

- Last iterate of the constant step-size SGD may not converge, Why ?

  Noise = Additive (model noise) + Multiplicative (SGD sampling noise)

  Additive noise forces to use **variance reduction techniques** for SGD to converge.

Last Iterate of SGD

- Last iterate of the constant step-size SGD may not converge, Why ?

  Noise = Additive (model noise) + Multiplicative (SGD sampling noise)

  Additive noise forces to use **variance reduction techniques** for SGD to converge.
- The noiseless setting: We make the hypothesis that the model is perfect, i.e., there is no additive noise, i.e there exists a perfect regressor θ_*

$$\langle \theta_*, x \rangle = y \quad a.s.$$

Last iterate of SGD should converge in this model !

Noiseless Least Squares

- Merits of the noiseless setting. Captures the modern machine learning architecture: **overparameterization** and **interpolation** (w.r.t. training loss)

Noiseless Least Squares

- Merits of the noiseless setting. Captures the modern machine learning architecture: **overparameterization** and **interpolation** (w.r.t. training loss)
- Non-strongly convex. For, strongly convex we have linear rates on last iterate. However, for non-strongly convex it was **open**.

Noiseless Least Squares

- Merits of the noiseless setting. Captures the modern machine learning architecture: **overparameterization** and **interpolation** (w.r.t. training loss)
- Non-strongly convex. For, strongly convex we have linear rates on last iterate. However, for non-strongly convex it was **open**.
- Covariance The covariance operator on $\mathcal{H}$:

$$\mathbf{H} := \mathbb{E}_\rho[x \otimes x].$$

The non-strongly convex setting corresponds to the **smallest eigen value** being **arbitrarily small** and close to 0.

Main Result

Recall, **Risk** : $\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_\rho \left(\langle \theta, x \rangle - Y\right)^2$, **SGD**: $\theta_{t+1} = \theta_t - \gamma \left(\langle \theta_t, x_t \rangle - y_t\right) x_t$.

Main Result

Recall, **Risk** : $\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_\rho\left(\langle\theta, x\rangle - Y\right)^2$, **SGD**: $\theta_{t+1} = \theta_t - \gamma\left(\langle\theta_t, x_t\rangle - y_t\right)x_t$.

$$\exists \mathrm{R} \;\; s.t.\; \mathbb{E}\left[\|x\|^2 \; xx^\top\right] \preccurlyeq \mathrm{R}\mathbf{H} \quad and \quad \|\theta_*\|_{\mathcal{H}} < +\infty \tag{1}$$

Main Result

Recall, **Risk** : $\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}_\rho\left(\langle\theta, x\rangle - Y\right)^2$, **SGD**: $\theta_{t+1} = \theta_t - \gamma\left(\langle\theta_t, x_t\rangle - y_t\right)x_t$.

$$\exists \mathrm{R} \;\; s.t. \; \mathbb{E}\left[\|x\|^2 \; xx^\top\right] \preccurlyq \mathrm{R}\mathbf{H} \quad and \quad \|\theta_*\|_{\mathcal{H}} < +\infty \tag{1}$$

Main Result

For $T \geqslant 2$, if we set $\gamma = (4R\ln(T))^{-1}$, we have the following bound for the expected risk of the estimator given by the T^{th} iterate of SGD:

$$\mathbb{E}\mathcal{R}(\theta_T) \leqslant 3 \; R \; \|\theta_*\|_{\mathcal{H}}^2 \; \frac{\ln(T)}{T}. \tag{2}$$

Non-parametric Rates

With further refinements over the **spectrum of co-variance i.e. capacity condition**

$$\exists \ \alpha > 0, R_\alpha > 0 \ s.t. \ \mathbb{E}\left[\langle x, \mathbf{H}^{-\alpha} x\rangle xx^\top\right] \preccurlyeq R_\alpha \mathbf{H} \tag{3}$$

Non-parametric Rates

With further refinements over the **spectrum of co-variance i.e. capacity condition**

$$\exists\ \alpha > 0, R_\alpha > 0 \ s.t.\ \mathbb{E}\left[\left\langle x, \mathbf{H}^{-\alpha} x\right\rangle x x^\top\right] \preccurlyeq R_\alpha \mathbf{H} \tag{3}$$

and **regularity of optimum i.e. source condition** like

$$\exists\ \beta > -1, C_\beta > 0 \ s.t.\ C_\beta = \|\mathbf{H}^{-\beta/2}\theta_*\|_{\mathcal{H}}^2 \tag{4}$$

Non-parametric rates

For $T \geqslant 3$, where $\gamma^{1-\alpha} \leqslant (32\xi_\alpha R_\alpha)^{-1}$ and $\xi_\alpha = \sum_{n\geqslant 1} \frac{1}{n^{1+\alpha}}$, we have

$$\mathbb{E}\mathcal{R}(\theta_T) \leqslant 2\left(\frac{1+\beta}{\gamma}\right)^{1+\beta} \frac{C_\beta}{T^{1+\alpha\wedge\beta}} \tag{5}$$

Conclusion

Contributions:

- **No additive noise** implies **no variance reduction**. Last iterate of the **constant step-size SGD** converges!

Perspectives:

Conclusion

Contributions:

- **No additive noise** implies **no variance reduction**. Last iterate of the **constant step-size SGD** converges!
- A new **Lyapunov technique** to control the bias error in standard least square analysis.

Perspectives:

Conclusion

Contributions:

- **No additive noise** implies **no variance reduction**. Last iterate of the **constant step-size SGD** converges!
- A new **Lyapunov technique** to control the bias error in standard least square analysis.

Perspectives:

- Insights into optimization of general convex (even non-convex) overparamaterized models.

Conclusion

Contributions:

- **No additive noise** implies **no variance reduction**. Last iterate of the **constant step-size SGD** converges!
- A new **Lyapunov technique** to control the bias error in standard least square analysis.

Perspectives:

- Insights into optimization of general convex (even non-convex) overparamaterized models.
- A simple and effective setting for understanding interplay between **momentum** with **stochastic/multiplicative** noise.