Fast and Faster Convergence of SGD for Over-Parameterized Models

Sharan Vaswani [1] Francis Bach [2] Mark Schmidt [1]
[1] University of British Columbia [2] INRIA, ENS, PSL Research University

Contributions

- We prove that, under a strong growth condition on the stochastic gradients, SGD with Nesterov momentum attains the accelerated convergence rate of the deterministic setting.
- Under this growth condition, we prove that SGD converges as fast as full-batch gradient descent for (strongly)-convex and non-convex functions.
- We show that a weaker growth condition is satisfied for smooth, convex losses for over-parametrized models that interpolate the data.
- We show that these results lead to a modified perceptron algorithm that has an accelerated rate of decrease on the number of mistakes.

General Setup

Objective: Find $w^* \in \arg\min f(w)$ assuming access to unbiased noisy gradients $\nabla f(w,z)$ such that $\mathbb{E}_z[\nabla f(w,z)] = \nabla f(w)$. Assumptions on $f(x)$:

- L-smoothness and μ-strong convexity.
- **Strong Growth Condition (SGC)**: $\mathbb{E}_z \|\nabla f(w,z)\|^2 \leq \rho \|\nabla f(w)\|^2$.
- Important special case: Finite sums: $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$.
- SGC $\implies \mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2$.
- **Interpolation**: $\nabla f_i(w^*) = 0$.

Algorithms

- **Constant step-size Stochastic Gradient Descent (SGD)**:

$$w_{k+1} = w_k - \eta \nabla f(w_k, z_k) \tag{1}$$

- **Constant step-size SGD with Nesterov acceleration**:

$$w_{k+1} = \zeta_k - \eta \nabla f(\zeta_k, z_k) \tag{2}$$
$$\zeta_k = \alpha_k v_k + (1-\alpha_k) w_k \tag{3}$$
$$v_{k+1} = \beta_k v_k + (1-\beta_k)\zeta_k - \gamma_k \eta \nabla f(\zeta_k, z_k). \tag{4}$$

Convergence of constant step-size SGD with Nesterov acceleration

Theorem (Strongly convex)

Under L-smoothness, μ-strong-convexity, ρ-SGC, SGD with Nesterov acceleration with

$$\gamma_k = \frac{1}{\sqrt{\mu\eta\rho}} \quad ; \quad \beta_k = 1 - \sqrt{\frac{\mu\eta}{\rho}} \quad ; \quad b_{k+1} = \frac{\sqrt{\mu}}{\left(1-\sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$

$$a_{k+1} = \frac{1}{\left(1-\sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \quad ; \quad \alpha_k = \frac{\gamma_k\beta_k b_{k+1}^2\eta}{\gamma_k\beta_k b_{k+1}^2\eta + a_k^2} \quad ; \quad \eta = \frac{1}{\rho L},$$

results in the following convergence rate:

$$\mathbb{E} f(w_{k+1}) - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho^2 L}}\right)^k \left[f(w_0) - f(w^*) + \frac{\mu}{2}\|w_0 - w^*\|^2\right].$$

Theorem (Convex)

Under L-smoothness, convexity, ρ-SGC, SGD with Nesterov acceleration with

$$\gamma_k = \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{k-1}^2}}{2} \quad ; \quad a_{k+1} = \gamma_k\sqrt{\eta\rho} \quad ; \quad \alpha_k = \frac{\gamma_k\eta}{\gamma_k\eta + a_k^2} \quad ; \quad \eta = \frac{1}{\rho L},$$

results in the following convergence rate:

$$\mathbb{E} f(w_{k+1}) - f(w^*) \leq \frac{2\rho^2 L}{k^2}\|w_0 - w^*\|^2.$$

- First result showing that SGD with Nesterov momentum matches the rates of the deterministic accelerated method.

Convergence of constant step-size SGD

- Constant step-size SGD matches the deterministic rates of convergence for (strongly)-convex functions (Schmidt, Le Roux '13).

Theorem (Non-Convex)

Under L-smoothness, ρ-SGC, SGD with a constant step-size $\eta = \frac{1}{\rho L}$ attains the following convergence rate:

$$\min_{i=0,1,\ldots k-1} \mathbb{E}\left[\|\nabla f(w_i)\|^2\right] \leq \left(\frac{2\rho L}{k}\right)[f(w_0) - f^*].$$

- First result for non-convex functions under interpolation-like conditions.

Theorem (Non-Convex + PL)

Under L-smoothness, ρ-SGC and if f satisfies the Polyak- Lojasiewicz inequality with constant μ, then SGD with a constant step-size $\eta = \frac{1}{\rho L}$ attains the following convergence rate:

$$\mathbb{E}[f(w_{k+1}) - f^*] \leq \left(1 - \frac{\mu}{\rho L}\right)^k [f(w_0) - f^*].$$

- Under specific conditions, the PL inequality is satisfied for non-convex functions occurring in neural networks, matrix completion and phase retrieval.

Relaxing the assumptions

- **Weak growth condition (WGC)**:

$$\mathbb{E}_z\|\nabla f(w,z)\|^2 \leq 2\rho L[f(w) - f(w^*)]. \tag{5}$$

Equivalently, in the finite-sum setting,

$$\mathbb{E}_i\|\nabla f_i(w)\|^2 \leq 2\rho L[f(w) - f(w^*)]. \tag{6}$$

- **Relation between the WGC and SGC**:

$$L\text{-smoothness, } \rho\text{-WGC, } \mu\text{-PL} \implies \frac{\rho L}{\mu}\text{-SGC}$$
$$L\text{-smoothness, } \rho\text{-SGC,} \implies \rho\text{-WGC}$$

Convergence of constant step-size SGD under the WGC

Theorem (Strongly-convex)

Under L-smoothness, μ-strong-convexity, ρ-WGC, SGD with a constant step-size $\eta = \frac{1}{\rho L}$ achieves the following rate:

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \leq \left(1 - \frac{\mu}{\rho L}\right)^k \|w_0 - w^*\|^2.$$

Theorem (Convex)

Under L-smoothness, convexity, ρ-WGC, SGD with a constant step-size $\eta = \frac{1}{4\rho L}$ and iterate averaging achieves the following rate:

$$\mathbb{E}[f(\bar{w}_k)] - f(w^*) \leq \frac{4L(1+\rho)\|w_0 - w^*\|^2}{k}.$$

Here, $\bar{w}_k = \frac{\left[\sum_{i=1}^k w_i\right]}{k}$ is the averaged iterate after k iterations.

Growth conditions in practice

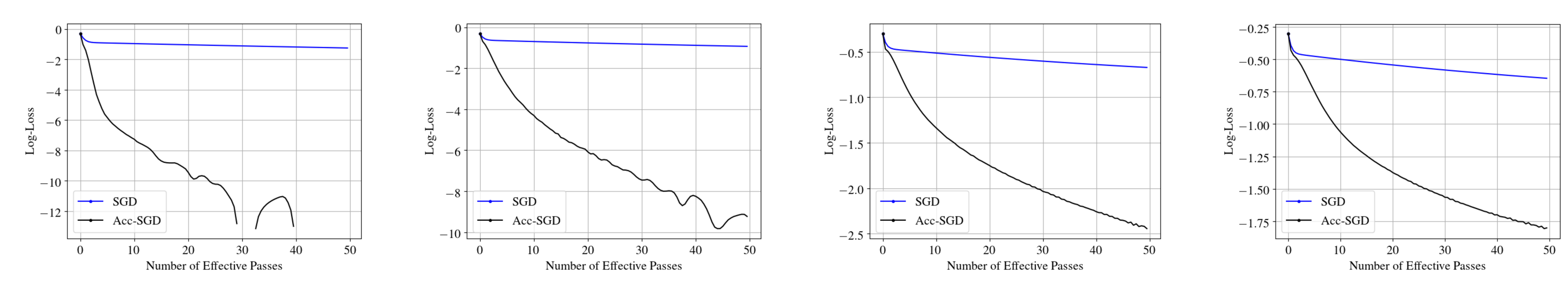
Proposition

If the function $f(\cdot)$ is convex and has a finite-sum structure for a model that interpolates the data and L_{max} is the maximum smoothness constant amongst the functions $f_i(\cdot)$, then for all w, $\mathbb{E}_i\|\nabla f_i(w)\|^2 \leq 2L_{max}[f(w) - f(w^*)]$

Accelerated perceptron using squared-hinge loss:

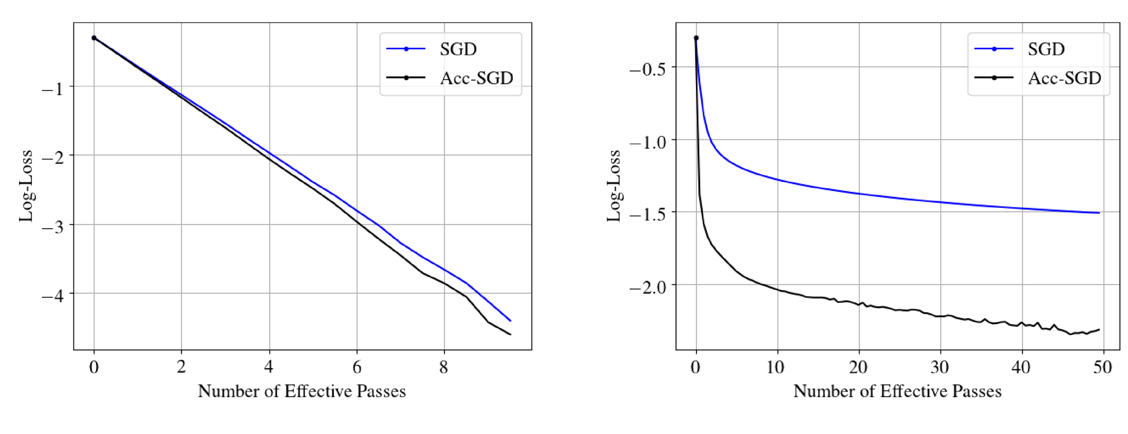
- For linearly separable data with margin τ and a finite support of size c, the squared-hinge loss satisfies the SGC with the constant $\rho = \frac{c}{\tau^2}$.
- If $f(w,x,y)$ represents the loss on the point (x,y) and $\mathbb{P}(yx^\top w_k \geqslant 0)$ is the number of mistakes made by the algorithm after k iterations, then $\mathbb{P}(yx^\top w \leqslant 0) \leqslant \mathbb{E}_{x,y} f(w,x,y)$.
- Above lemmas + Theorem 2 $\implies O\left(\frac{1}{\tau^6 k^2}\right)$ mistake-bound while only requiring one gradient per iteration.

Experiments



(a) $\tau = 0.1$ (b) $\tau = 0.05$ (c) $\tau = 0.01$ (d) $\tau = 0.005$

Figure: Comparison of SGD and variants of accelerated SGD on a synthetic linearly separable dataset with margin τ. Accelerated SGD with $\eta = \tau/L$ leads to faster convergence as compared to SGD with $\eta = 1/L$.



(a) CovType (b) Protein

Figure: Comparison of SGD and accelerated SGD for learning a linear classifier with RBF features on the (a) CovType and (b) Protein datasets. Accelerated SGD leads to better performance as compared to SGD with $\eta = 1/L$.

- Can use the line-search procedure in (Schmidt, Le Roux, Bach'13) to obtain better convergence in practice.

Related Work

- Schmidt, Le Roux'13: "Fast convergence of stochastic gradient descent under a strong growth condition."
- Cevher, Vu'18: "On the linear convergence of the stochastic gradient method with constant step-size."
- Ma, Bassily, Belkin'18: "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning."
- Liu, Belkin'18: "Mass: an accelerated stochastic method for over-parametrized learning."