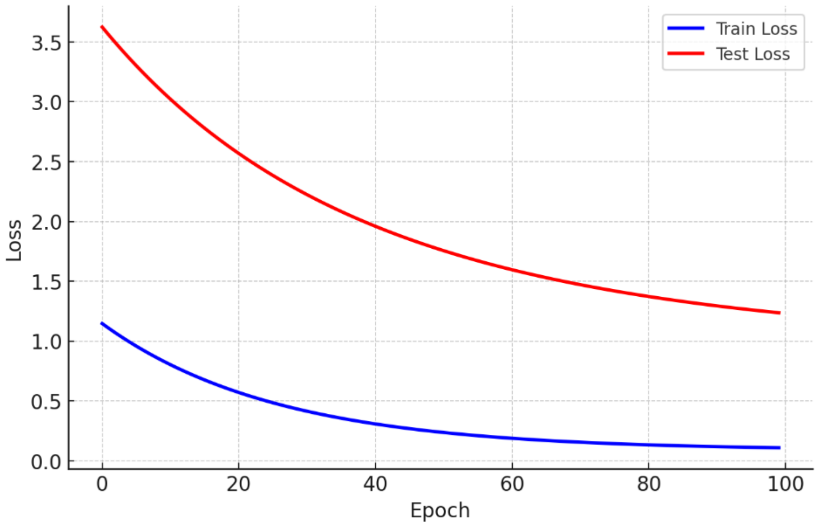
Stability and Generalization for Stochastic Recursive Momentum-based Algorithms for (Strongly-)Convex One to K-Level Stochastic Optimizations

Xiaokang Pan[1], Xingyu Li[2] , Jin Liu[1], Tao Sun[3], Kai Sun[4], Lixing Chen[5], Zhe Qu[1]

[1] Central South University; [2] Tulane University ; [3] National University of Defense Technology

[4] Xi'an Jiaotong University; [5] Shanghai Jiao Tong University



In machina learning we are more interested in the prediction behavior from the perspective of learning, i.e., how these models would behave on testing examples, which is much less studied for stochastic optimization.

1. Formulation: $\mathbb{E}_{S,A}[F(A(S)) - F(x_*)]$

2. Decomposition :

$$\mathbb{E}_{S,A}[F(A(S)) - F_S(A(S))] + \mathbb{E}_{S,A}[F_S(A(S)) - F_S(x_*^S)]$$

We call the first term mentioned above the generalization error; it quantifies the shift in model performance from training to testing. The second term is known as the optimization error, which measures the algorithm's effectiveness in minimizing empirical risk.

Questions

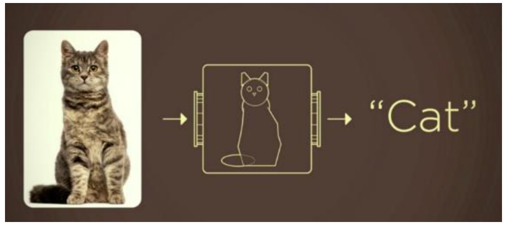
1. One-level : $\min_{x\in\mathcal{X}}\Big\{F(x)=\mathbb{E}_\nu[f_\nu(x)]\Big\}$

2. Two-level : $\min_{x\in\mathcal{X}}\Big\{F(x)=f\circ g(x)=\mathbb{E}_\nu[f_\nu(\mathbb{E}_\omega[g_\omega(x)])]\Big\}$

3. Multi-level : $\min_{x\in\mathcal{X}}\Big\{F(x)=f_K\circ f_{K-1}\circ\cdots\circ f_1(x)=\mathbb{E}_{\nu^{(K)}}[f_K^{\nu^{(K)}}(\cdots\mathbb{E}_{\nu^{(1)}}[f_1^{\nu^{(1)}}(x)])]\Big\}$

Applications

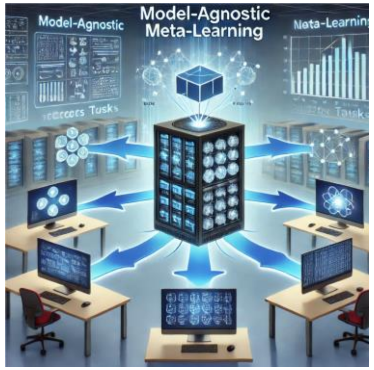
1. One-level : Image classification, Style transfer.



2. Two-level : Risk averse portfolio

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}_t^\top\boldsymbol{\theta} - \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{r}_t^\top\boldsymbol{\theta} - \frac{1}{T}\sum_{j=1}^{T} \mathbf{r}_j^\top\boldsymbol{\theta}\right)^2.$$

3. Multi-level :



$$\min_{\mathbf{x}} F(\boldsymbol{\theta}) := \frac{1}{M}\sum_{m=1}^{M} F_m\left(\mathbf{x} - \alpha\nabla F_m(\mathbf{x})\right),$$

with $F_m(\boldsymbol{\theta}) := \mathbb{E}_{\xi_m}\left[f\left(\boldsymbol{\theta};\xi_m\right)\right]$

Our Contributions:

1. We define uniform stability for K-level optimizations and link it to generalization error.

2. Our theoretical results indicate that fewer iterations and proper step sizes will improve algorithm stability of stability in the convex setting.

3. Increasing levels impact the generalization error by reducing stability and increasing the variance between the combined stochastic gradient and the true gradient.

4. Additionally, using more samples initially can enhance performance, thereby improving the efficiency of STORM-based algorithms.

Uniform Stability

1. One-level : A is uniformly stable if $\forall i \in [1, n]$

$$\mathbb{E}_A[\|A(S) - A(S^i)\|] \leq \epsilon$$

2. Two-level : A is uniformly stable if $\forall i \in [1, n]$ and $\forall j \in [1, m]$

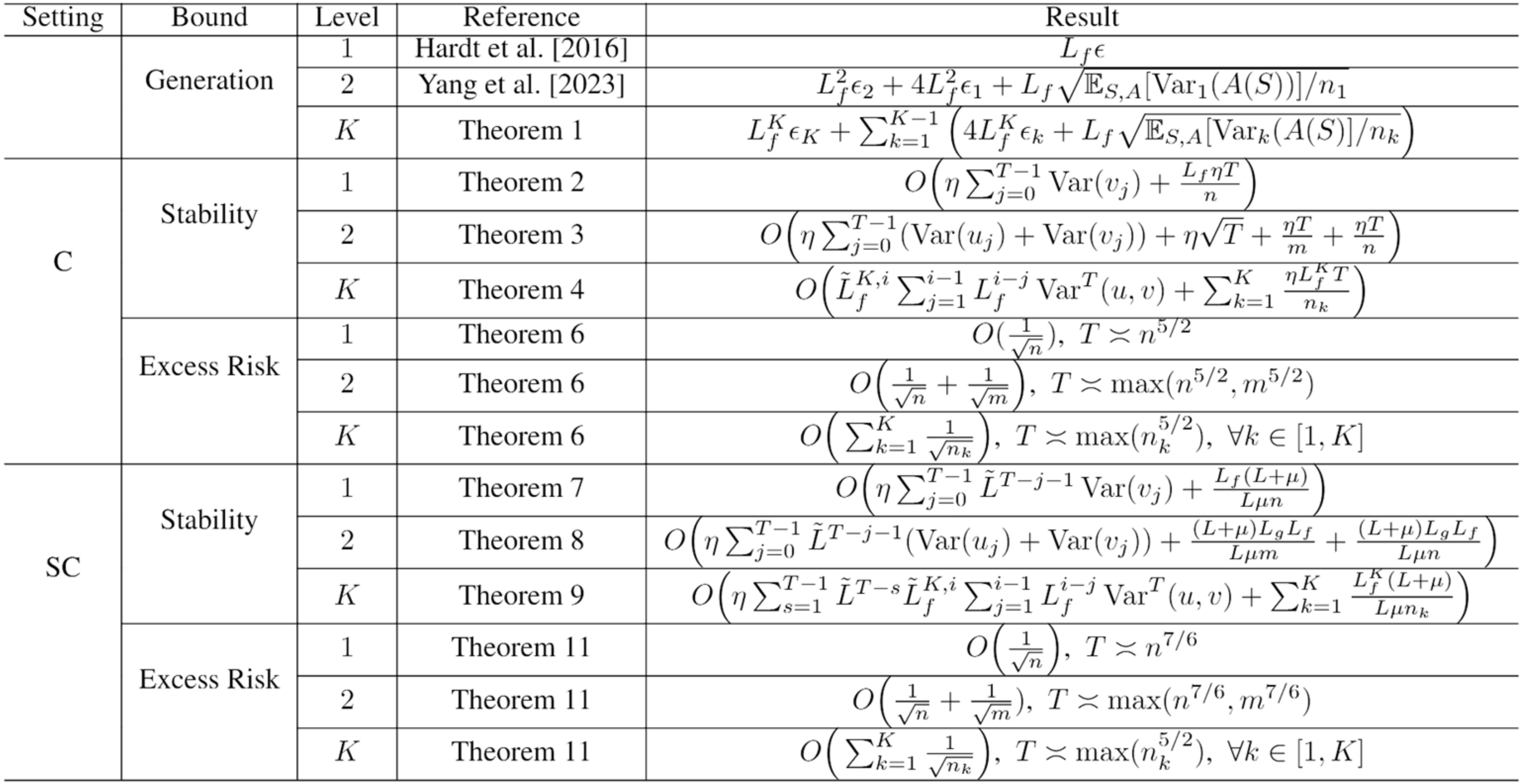
$$\mathbb{E}_A[\|A(S) - A(S^{i,\nu})\|] \leq \epsilon_\nu \quad \text{and} \quad \mathbb{E}_A[\|A(S) - A(S^{j,\omega})\|] \leq \epsilon_\omega$$

3. Multi-level : A is uniformly stable if $\forall k \in [1, K]$ and $\forall l \in [1, n_k]$

$$\mathbb{E}_A[\|A(S) - A(S^{l,k})\|] \leq \epsilon_k$$

Results

Table 1: Summary of our results.



Setting	Bound	Level	Reference	Result
	Generation	1	Hardt et al. [2016]	$L_f \epsilon$
		2	Yang et al. [2023]	$L_f^2 \epsilon_2 + 4L_f^2 \epsilon_1 + L_f \sqrt{\mathbb{E}_{S,A}[\mathrm{Var}_1(A(S))]/n_1}$
		K	Theorem 1	$L_f^K \epsilon_K + \sum_{k=1}^{K-1} \left(4L_f^K \epsilon_k + L_f \sqrt{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]/n_k}\right)$
C	Stability	1	Theorem 2	$O\left(\eta \sum_{j=0}^{T-1} \mathrm{Var}(v_j) + \frac{L_f \eta T}{n}\right)$
		2	Theorem 3	$O\left(\eta \sum_{j=0}^{T-1} (\mathrm{Var}(u_j) + \mathrm{Var}(v_j)) + \eta\sqrt{T} + \frac{\eta T}{m} + \frac{\eta T}{n}\right)$
		K	Theorem 4	$O\left(\tilde{L}_f^{K,i} \sum_{j=1}^{i-1} L_f^{i-j} \mathrm{Var}^T(u,v) + \sum_{k=1}^{K} \frac{\eta L_f^K T}{n_k}\right)$
	Excess Risk	1	Theorem 6	$O(\frac{1}{\sqrt{n}}),\ T \asymp n^{5/2}$
		2	Theorem 6	$O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right),\ T \asymp \max(n^{5/2}, m^{5/2})$
		K	Theorem 6	$O\left(\sum_{k=1}^{K} \frac{1}{\sqrt{n_k}}\right),\ T \asymp \max(n_k^{5/2}),\ \forall k \in [1, K]$
SC	Stability	1	Theorem 7	$O\left(\eta \sum_{j=0}^{T-1} \tilde{L}^{T-j-1} \mathrm{Var}(v_j) + \frac{L_f(L+\mu)}{L\mu n}\right)$
		2	Theorem 8	$O\left(\eta \sum_{j=0}^{T-1} \tilde{L}^{T-j-1} (\mathrm{Var}(u_j) + \mathrm{Var}(v_j)) + \frac{(L+\mu)L_g L_f}{L\mu m} + \frac{(L+\mu)L_g L_f}{L\mu n}\right)$
		K	Theorem 9	$O\left(\eta \sum_{s=1}^{T-1} \tilde{L}^{T-s} \tilde{L}_f^{K,i} \sum_{j=1}^{i-1} L_f^{i-j} \mathrm{Var}^T(u,v) + \sum_{k=1}^{K} \frac{L_f^K (L+\mu)}{L\mu n_k}\right)$
	Excess Risk	1	Theorem 11	$O\left(\frac{1}{\sqrt{n}}\right),\ T \asymp n^{7/6}$
		2	Theorem 11	$O\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}),\ T \asymp \max(n^{7/6}, m^{7/6})$
		K	Theorem 11	$O\left(\sum_{k=1}^{K} \frac{1}{\sqrt{n_k}}\right),\ T \asymp \max(n_k^{5/2}),\ \forall k \in [1, K]$

Main results

- Theorem 1 (Quantitative relationship between generalization and stability):

$$L_f^K \epsilon_K + \sum_{k=1}^{K-1} \Big(4L_f^K \epsilon_k + L_f \sqrt{\frac{\mathbb{E}_{S,A}[\mathrm{Var}_k(A(S)]}{n_k}} \Big),$$

- Theorem 6 (Excess Risk Bounds of three Algorithms):

1. Setting T $= O(n^{2.5}), \eta = O(T^{-0.8})$ and $\beta = O(T^{-0.8})$ we can obtain

$$\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\frac{1}{\sqrt{n}}\Big)$$

2. Setting T $= O(\max(n^{2.5}, m^{2.5})), \eta = O(T^{-0.8})$ and $\beta = O(T^{-0.8})$ we can obtain

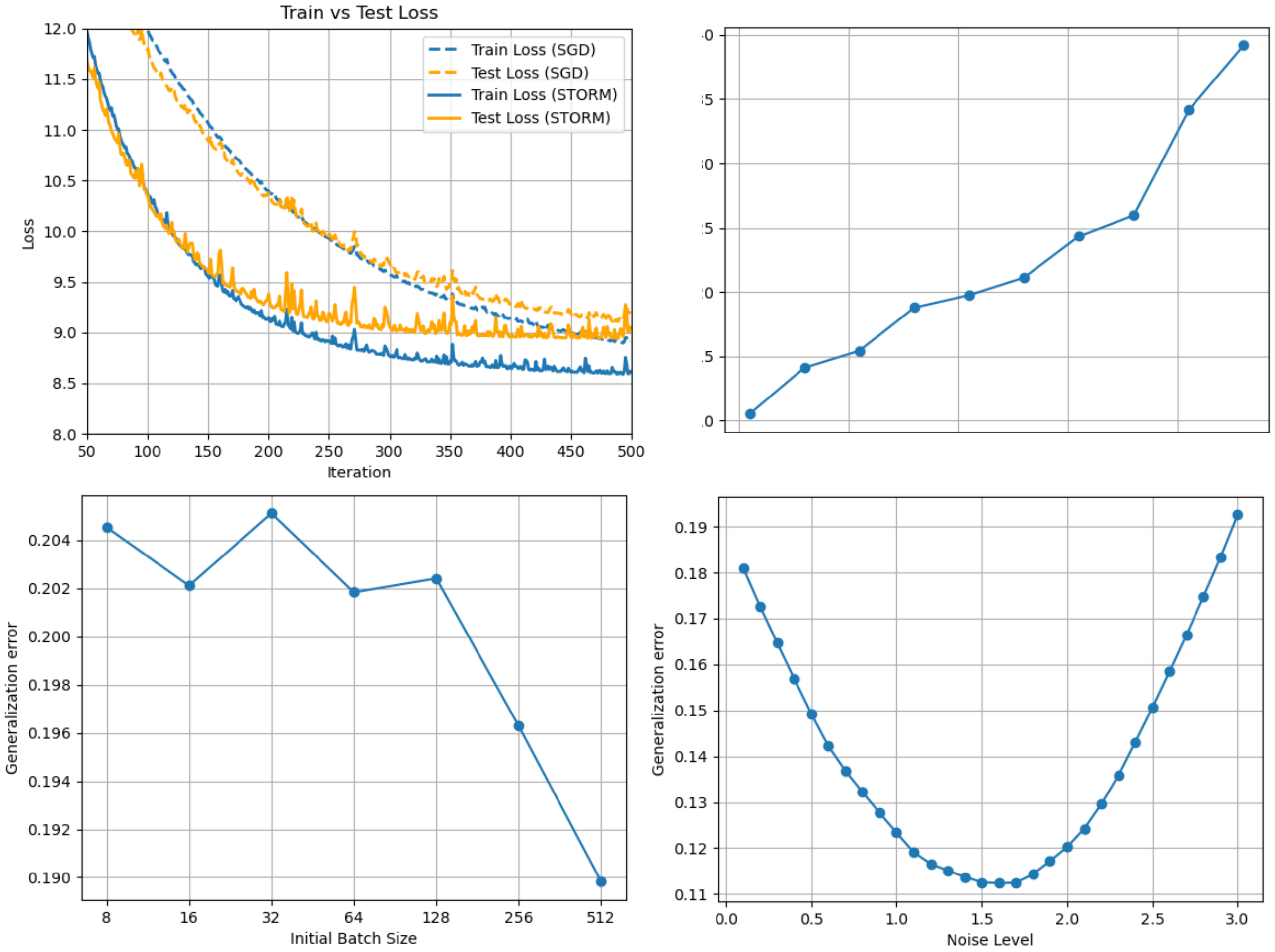
$$\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)$$

3. Setting T $= O\big(n_k^{2.5}\big), \eta = O(T^{-0.8})$ and $\beta = O(T^{-0.8})$ for any $k \in [1, K]$we can obtain

$$\mathbb{E}_{S,A}[F(A(S)) - F(x_*)] = O\Big(\sum_{k=1}^{K} \frac{1}{\sqrt{n_k}}\Big)$$

Experiments Results

Figure 1: Experiments of our results.



- First, we examined the performance of STORM versus SGD in fitting a univariate quintic polynomial.

- Second, we investigated how varying the number of levels affects generalization error within a two-level optimization framework.

- Third, we explored the impact of the initial iteration batch size on generalization.

- Fourth, we investigated the impact of noise on generalization.

THANKS FOR ALL