# Black-Box Generalization: Stability of Zeroth-Order Learning

Konstantinos Nikolakakis, Farzin Haddadpour, Dionysios Kalogerias, Amin Karbasi

Yale University

## Motivation — Applications

- The loss function is unknown, but a limited amount of functions evaluations is available
- Optimization error: Stochastic differences of function evaluations provide optimal approximations of the gradient (Duchi et al.)
- Black-box adversarial attacks, federated learning, reinforcement learning, what about generalization?

## Assumptions

- The loss function is $L$-Lipschitz

$$\|f(w,z)-f(u,z)\|_2 \le L\|w-u\|_2, \quad \forall z \in \mathcal{Z}.$$

- The gradient of the loss function is $\beta$-Lipschitz

$$\|\nabla_w f(w,z)-\nabla_u f(u,z)\|_2 \le \beta\|w-u\|_2, \quad \forall z \in \mathcal{Z}.$$

## Lemma (Growth Recursion: ZoSS)

*Consider the sequences of updates* $\{\tilde G_t\}_{t=1}^T$, $\{\tilde G'_t\}_{t=1}^T$, *define* $\Delta\tilde G_t \triangleq \mathbb{E}[\|\tilde G_t(w_t)-\tilde G'_t(w'_t)\|]$ *and* $\Gamma_K^d \triangleq \sqrt{(3d-1)/K}+1$. *Let* $w_0=w'_0$ *be the starting point,* $w_{t+1}=\tilde G_t(w_t)$ *and* $w'_{t+1}=\tilde G'_t(w'_t)$ *for any* $t\in\{1,\dots,T\}$. *Then for any* $w_t, w'_t \in \mathbb{R}^d$

$$\Delta\tilde G_t \le \begin{cases} \left(1+\alpha_t\beta\Gamma_K^d\right)\|w_t-w'_t\| + \mu\beta\alpha_t(3+d)^{3/2}, & \tilde G_t(\cdot)=\tilde G'_t(\cdot),\\ \|w_t-w'_t\| + 2\alpha_t L\Gamma_K^d + \mu\beta\alpha_t(3+d)^{3/2}, & \tilde G_t(\cdot)\neq\tilde G'_t(\cdot).\end{cases}$$

## Lemma (Growth Recursion: Mini-Batch ZoSS)

*Consider the sequences of updates* $\{\tilde G_{J_t}\}_{t=1}^T$ *and* $\{\tilde G'_{J_t}\}_{t=1}^T$, *define* $\Delta\tilde G_{J_t} \triangleq \mathbb{E}[\|\tilde G_{J_t}(w_t)-\tilde G'_{J_t}(w'_t)\|]$ *and* $\mu \le cL\Gamma_K^d/(n\beta(3+d)^{3/2})$. *Let* $w_0=w'_0$ *be the starting point,* $w_{t+1}=\tilde G_{J_t}(w_t)$ *and* $w'_{t+1}=\tilde G'_{J_t}(w'_t)$ *for any* $t\in\{1,\dots,T\}$. *Then for any* $w_t, w'_t\in\mathbb{R}^d$ *and* $t\ge 0$

$$\Delta\tilde G_{J_t} \le \begin{cases} \left(1+\beta\alpha_t\Gamma_K^d\right)\|w_t-w'_t\| + \frac{cL\alpha_t}{n}\Gamma_K^d, & \tilde G_{J_t}(\cdot)=\tilde G'_{J_t}(\cdot)\\ \left(1+\frac{m-1}{m}\beta\alpha_t\Gamma_K^d\right)\|w_t-w'_t\| + \left(\frac{2L\alpha_t}{m}+\frac{cL\alpha_t}{n}\right)\Gamma_K^d, & \tilde G_{J_t}(\cdot)\neq\tilde G'_{J_t}(\cdot).\end{cases}$$

## Lemma (ZoSS Stability — Nonconvex Loss)

*Consider the ZoSS algorithm with final-iterate estimates* $A(S)$ *and* $A(S')$, *corresponding to the data-sets* $S, S'$, *respectively (that differ in exactly one entry). Then the discrepancy* $\delta_T \triangleq \|A(S)-A(S')\|$, *under the event* $\mathcal{E}_{\delta_{t_0}}$, *satisfies the inequality*

$$\mathbb{E}[\delta_T|\mathcal{E}_{\delta_{t_0}}] \le \left(\frac{2L}{n}\Gamma_K^d + \mu\beta(3+d)^{3/2}\right)\sum_{t=t_0+1}^{T}\alpha_t\prod_{j=t+1}^{T}\left(1+\beta\alpha_j\Gamma_K^d\right).$$

## Problem Statement

- Let $f(w,z)$ be the loss at $w\in\mathbb{R}^d$ for some example $z\in\mathcal{Z}$.
- Given a dataset $S\triangleq\{z_i\}_{i=1}^n$ of i.i.d $z_i\sim\mathcal{D}$
- Find the parameters $w^*$ such that $w^*\in\arg\min_w R(w)$, where $R(w)\triangleq\mathbb{E}_{Z\sim\mathcal{D}}[f(w,Z)]$

Since $\mathcal{D}$ is not known, we consider the empirical risk

$$R_S(w)\triangleq\frac{1}{n}\sum_{i=1}^n f(w;z_i),$$

and the corresponding empirical risk minimization (ERM) problem

- Find $W_S^*\in\arg\min_w R_S(w)$
- $\epsilon_{\text{gen}}\triangleq\mathbb{E}[R(A(S))-R_S(A(S))]$ (Generalization Error)
- $\epsilon_{\text{excess}}\triangleq\mathbb{E}_{S,A}[R(A(S))]-R(w^*)=\underbrace{\mathbb{E}_{S,A}[R(A(S))-R_S(A(S))]}_{\epsilon_{\text{gen}}}+\underbrace{(\mathbb{E}_{S,A}[R_S(A(S))]-R(w^*))}_{\epsilon_{\text{opt}}}.$

For i.i.d. $S,S'\in\mathcal{Z}^n$ that differ in one entry, $\sup_z\mathbb{E}_A[f(A(S),z)-f(A(S'),z)]\le\epsilon_{\text{stab}}$, for some $\epsilon_{\text{stab}}>0$, then $\epsilon_{\text{gen}}\le\epsilon_{\text{stab}}$ and $\epsilon_{\text{stab}}\le L\sup_{S,S'}\mathbb{E}_A\|A(S)-A(S')\|$.

## Zeroth-Order Stochastic Search (ZoSS)

As a *gradient-free* alternative of the classical SGD algorithm, we consider the ZoSS scheme, with (single-example update) update rule

$$\Delta f^{K,\mu}_{w,z_{i_t}}\equiv\Delta f^{K,\mu,\mathbf{U}^t}_{w,z_{i_t}}\triangleq\frac{1}{K}\sum_{k=1}^K\frac{f(w+\mu U_k^t,z_{i_t})-f(w,z_{i_t})}{\mu}U_k^t,$$

$$W_{t+1}=W_t-\alpha_t\Delta f^{K,\mu}_{W_t,z_{i_t}},\quad U_k^t\sim\mathcal{N}(0,I_d),\quad\mu\in\mathbb{R}^+,$$

where $\alpha_t\ge0$ is the learning rate. At every iteration $t$, ZoSS generates $K$ i.i.d. standard normal random vectors $U_k^t, k=1,\dots,K$, and obtains $K+1$ loss evaluations on perturbed model inputs.

## Error Decomposition

The stability error of ZoSS at time $t$ breaks down into the stability error of SGD and an approximation error due to missing gradient information. Let $G_t(\cdot)$ and $G'_t(\cdot)$ be SGD iterations

$$G_t(w)\triangleq w-\alpha_t\nabla f(w,z_{i_t}),\quad G'_t(w)\triangleq w-\alpha_t\nabla f(w,z'_{i_t})$$

under inputs $S,S'$ respectively, and let $i_t\in\{1,2,\dots,n\}$ be a random index chosen uniformly and independently by the random selection rule of the algorithm, for all $t\le T$. Similarly we use the notation $\tilde G(\cdot)$ and $\tilde G'(\cdot)$ to denote the iteration mappings of ZoSS, i.e.,

$$\tilde G_t(w)\triangleq w-\alpha_t\Delta f_{w,z_{i_t}},\quad\tilde G'_t(w)\triangleq w-\alpha_t\Delta f_{w,z'_{i_t}}.$$

Then the iterate stability error $\tilde G_t(w)-\tilde G'_t(w')$ of ZoSS, for any $w,w'\in\mathbb{R}^d$ and for all at $t\le T$, may be decomposed as

$$\tilde G_t(w)-\tilde G'_t(w')\propto\underbrace{G_t(w)-G'_t(w')}_{\epsilon_{\text{GBstab}}}+\underbrace{\left[\nabla f(w,z_{i_t})-\Delta f_{w,z_{i_t}}\right]+\left[\nabla f(w',z'_{i_t})-\Delta f_{w',z'_{i_t}}\right]}_{\epsilon_{\text{est}}},$$

where $\epsilon_{\text{GBstab}}$ denotes the gradient-based stability error (associated with SGD), and $\epsilon_{\text{est}}$ denotes the gradient approximation error.

## Growth Recursion — Sketch of the Proof

Define $\mathbf{V}\triangleq\nabla f(w_t,z_{i_t})-\nabla f(w'_t,z_{i_t})$. We apply Taylor's expansion to find that for any $w_t,w'_t\in\mathbb{R}^d$ it is true that

$$\begin{aligned}\tilde G_t(w_t)-\tilde G'_t(w'_t)&=\tilde G_t(w_t)-\tilde G_t(w'_t)\\&=\underbrace{w_t-\alpha_t\nabla f(w_t,z_{i_t})}_{G(w_t)}-\underbrace{(w'_t-\alpha_t\nabla f(w'_t,z_{i_t}))}_{G'(w'_t)\equiv G(w'_t)}\\&-\alpha_t\left(\frac{1}{K}\sum_{k=1}^K\langle\mathbf{V},U_k^t\rangle U_k^t-\mathbf{V}\right)\\&-\frac{\alpha_t}{K}\sum_{k=1}^K\left(\frac{\mu}{2}(U_k^t)^\intercal\left(\nabla^2 f(W^*_{k,t},z_{i_t})-\nabla^2 f(W^\dagger_{k,t},z_{i_t})\right)U_k^t\right)U_k^t\end{aligned}$$

## Lemma (Variance Reduction)

*Let* $\mathbf{U}_k\in\mathbb{R}^d, k\in\{1,2\dots,K\}$ *be i.i.d standard Gaussian. Then for all* $\mathbf{V}\in\mathbb{R}^d$ *independent of all* $\mathbf{U}_k$

$$\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^K\langle\mathbf{V},\mathbf{U}_k\rangle\mathbf{U}_k-\mathbf{V}\right\|\middle|\mathbf{V}\right]\le\sqrt{\frac{3d-1}{K}}\|\mathbf{V}\|.$$

## Theorem (Nonconvex Bounded Loss)

*Consider the ZoSS algorithm with* $T$ *total number of iterates, stepsize* $\alpha_t\le C/t\Gamma_K^d$ $(C>0)$, *and fixed* $\mu\le cL\Gamma_K^d/n\beta(3+d)^{3/2}$ *for some* $c>0$. *Then*

$$|\epsilon_{\text{gen}}|\le\frac{\left(1+(C\beta)^{-1}\right)\left((2+c)CL^2\right)^{\frac{1}{C\beta+1}}}{n}(eT)^{\frac{C\beta}{C\beta+1}}.$$

## Summary of the Results

Generalization Error Bounds: ZoSS vs SGD

| Algorithm | Bound | NC | UB | MB |
|---|---|---|---|---|
| ZoSS (**this work**) $\alpha_t\le C/(t\Gamma_K^d)$ | $\frac{1+(C\beta)^{-1}}{n}\left((2+c)CL^2\right)^{\frac{1}{C\beta+1}}(eT)^{\frac{C\beta}{C\beta+1}}$ | ✓ | ✗ | ✗ |
| SGD, $\alpha_t\le C/t$ Hardt et al. [1] | $\frac{1+(C\beta)^{-1}}{n}\left(2CL^2\right)^{\frac{1}{C\beta+1}}(eT)^{\frac{C\beta}{C\beta+1}}$ | ✓ | ✗ | ✗ |
| ZoSS (**this work**) $\alpha_t\le C/t$ | $\frac{3e\left(1+(C\beta)^{-1}\right)^2}{2n}\left(1+(2+c)CL^2\right)T$ (*independent of both* d *and* $K$) | ✓ | ✗ | ✗ |
| ZoSS (**this work**) $\alpha_t\le\frac{\log\left(1+\frac{C\beta}{\Gamma_K^d}(\Gamma_K^d-1)\right)}{T\beta\sqrt{(3d-1)/K}}$ | $\frac{(2+c)CL^2}{n}$ | ✗ | ✓ | ✓ |
| SGD, $\alpha_t\le C/T$ Hardt et al. [1] | $\frac{2CL^2}{n}$ | ✗ | ✓ | ✓ |
| ZoSS (**this work**) $\alpha_t\le C/(T\Gamma_K^d)$ | $\frac{(2+c)L^2(e^{C\beta}-1)}{n\beta}$ | ✓ | ✓ | ✓ |
| ZoSS (**this work**) $\alpha_t\le\frac{\log(1+C\beta)}{T\beta\Gamma_K^d}$ | $\frac{(2+c)CL^2}{n}$ (proper choice of $C$ in previous bound) | ✓ | ✓ | ✓ |
| ZoSS (**this work**) $\alpha_t\le C/(t\Gamma_K^d)$ | $\frac{(2+c)L^2(eT)^{C\beta}}{n}\min\{C+\beta^{-1},C\log(T)\}$ | ✓ | ✓ | ✓ |