Hessian-Amended Random Perturbation (HARP) Using Noisy Zeroth-Order Oracle

Jingyi Zhu

MSML21: Mathematical and Scientific Machine Learning
Session 1: Optimization and Algorithms

August 16, 2021

Minimization Using *Few* Zeroth-Order Queries

$$\min_{\theta \in \mathbb{R}^d} L(\theta) \equiv \mathbb{E}_{\omega \sim \mathbb{P}}[\ell(\theta, \omega)]\,, \tag{1}$$

- **stochastic**: evaluation of $L(\theta)$ is corrupted by *noise*
- **limited-resource**: collecting $\ell(\cdot, \omega)$ is *expensive*

Stochastic Approximation (SA) Algorithms

$$\text{1st-order}: \hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k)\,. \tag{2}$$

w/ a_k is stepsize. We use gradient estimator using two ZO queries:

$$\hat{\boldsymbol{g}}_k(\hat{\theta}_k) = \frac{\ell(\hat{\theta}_k + c_k \boldsymbol{\Delta}_k, \omega_k^+) - \ell(\hat{\theta}_k - c_k \boldsymbol{\Delta}_k, \omega_k^-)}{2c_k} \boldsymbol{m}_k(\boldsymbol{\Delta}_k)\,. \tag{3}$$

w/ c_k is differencing magnitude, $\boldsymbol{\Delta}_k$ is perturbation, $\boldsymbol{m}_k(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^d$.

RDSA (random direction) [Erm69, Erm83]

$$\boldsymbol{\Delta}_k \sim \mathrm{Unif}(\mathbb{S}) \quad \text{and} \quad \boldsymbol{m}_k(\boldsymbol{\Delta}_k) = d\boldsymbol{\Delta}_k$$

SPSA (simultaneous perturbation) [Spa92]

$$\boldsymbol{\Delta}_k \sim [\mathrm{Unif}\{-1, 1\}]^d, \;\; \boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Delta}_k$$

SFSA (smoothed functional) [KO72]

$$\boldsymbol{\Delta}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \boldsymbol{\Delta}_k$$

Randomized FDSA (finite-difference)—effective as cyclic scheme

$$\boldsymbol{\Delta}_k \sim \mathrm{Unif}\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\} \quad \text{and} \quad \boldsymbol{m}_k(\boldsymbol{\Delta}_k) = d\boldsymbol{\Delta}_k$$

What if $\boldsymbol{\Delta}_k$ has zero-mean and $\boldsymbol{\Sigma}_k^{-1}$-covariance for $\boldsymbol{\Sigma}_k \succ \mathbf{0}$?

- Data-driven $\boldsymbol{\Sigma}_k$ can be $\mathcal{F}_k$-measurable random variable
- User-specified $\boldsymbol{\Sigma}_k$ can use domain knowledge

Scaling/Stepsize

Minimize $L(\theta) = 100\theta_1^2 + \theta_2^2$, with $\hat{\theta} = [1,1]^T$ and $c = 0.1$.

- SPSA generates $[1,1]^T$, $[1,-1]^T$, $[-1,1]^T$ and $[-1,-1]^T$ equally likely. $\mathbb{E}_{\boldsymbol{\Delta}}[\hat{\boldsymbol{g}}(\hat{\theta})]$ equals true gradient $\boldsymbol{g}(\hat{\theta}) = [100,1]^T$. But the variance is in the order of 10^4 assuming noise-free ZO queries.
- HARP draws $\boldsymbol{\Delta}$ from $\mathbf{0}$-mean and $\boldsymbol{\Sigma}^{-1}$-covariance distribution with $\boldsymbol{\Sigma} = \hat{\boldsymbol{H}}(\hat{\theta})$. Still unbiased, but covariance matrix norm is 2×10^2.

Correlation/Direction

- $\boldsymbol{\Sigma} = \boldsymbol{I}$ has *zero* off-diagonal elements implies that each component of $\boldsymbol{\Delta}$ is *independent*.
- Say $L(\theta) = 100\theta_1^2 + \theta_2^2 + \theta_1\theta_2$ with same $\hat{\theta}$ and c. The covariance magnitude of SPSA gradient estimator is around 4×10^4 and that of HARP is 8×10^2.

HARP handles scaling & correlation

$\boldsymbol{\Delta}_k$ follows a dist. w/ mean $\mathbf{0}$ and cov. $\hat{\boldsymbol{H}}_k^{-1}$, and $\boldsymbol{m}_k(\boldsymbol{\Delta}_k) = \hat{\boldsymbol{H}}_k \boldsymbol{\Delta}_k$

Theoretical Guarantee

Root-mean-squared error $\mathbb{E}[||\hat{\theta}_k - \theta^*||^2]^{1/2}$ is smaller when $\boldsymbol{\Sigma}_k = \hat{\boldsymbol{H}}_k$ than when $\boldsymbol{\Sigma}_k = \boldsymbol{I}$ for ill-conditioned problem.

- Expectation of estimation error goes to zero at the same[a] rate.
- Variance gets smaller while using $\boldsymbol{\Sigma}_k = \hat{\boldsymbol{H}}_k$.

[a] $k^{-1/3}$ when ω_k^+ and ω_k^- are independent and identically distributed (IID). $k^{-1/2}$ when $\omega_k^+ = \omega_k^-$, referred to as common random number (CRN).

Hessian can be estimated [Spa00, Spa09] provides principled way to estimate Hessian using *four* loss function evaluations. [Zhu21] proposes other forms that uses *two or more*.

However, computing issue persists, though [ZWS19] reduces per-iteration FLOPs from $O(d^3)$ to $O(d^2)$. Not comparable with $O(d)$ for generic first-order methods.

Feed Hessian estimate into both $\mathbf{\Sigma}_k$

1. **generate** $\mathbf{\Delta}_k$ so that it has a mean of $\mathbf{0}$ and a covariance $\hat{\boldsymbol{H}}_k^{-1}$
2. **collect** two noisy losses and **estimate** $\hat{g}_k$ using $\boldsymbol{m}_k(\mathbf{\Delta}_k)\left[\ell(\hat{\theta}_k + c_k\mathbf{\Delta}_k) - \ell(\hat{\theta}_k - c_k\mathbf{\Delta}_k)\right]/(2c_k)$
3. *may* **collect** additional queries and **estimate** $\hat{\boldsymbol{H}}_k$

Skew-quartic function is ill-conditioned: one large eigenvalues, and remaining eigenvalues are close to zero.

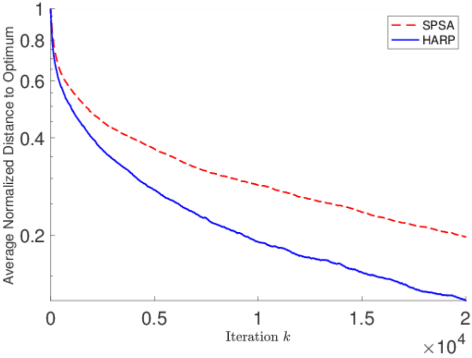

Figure: Performance of SPSA and HARP in terms of normalized distance $^{\|\hat{\theta}_k - \theta^*\|}/_{\|\hat{\theta}_0 - \theta^*\|}$ averaged across 25 independent replicates, and both algorithms use four ZO queries per iteration. The underlying loss function is the skew-quartic function with $d = 20$, and the noisy observation is corrupted by a $\mathcal{N}(0, 1)$ noise.

We consider generating adversarial perturbation universally for $I > 1$ images [CZS+17, CLC+18]:

$$\begin{cases} \min_{\theta} L(\theta) = \underbrace{\kappa||\theta||_2^2}_{\equiv L_1(\theta)} + \underbrace{\frac{1}{I}\sum_{i=1}^{I} \text{loss}(\zeta_i + \theta)}_{\equiv L_2(\theta)}, \\ \text{s.t. } (\zeta_i + \theta) \in [-0.5, 0.5]^d, \forall i, \end{cases} \tag{4}$$

where the constraint is to normalize the resulting pixels within $[-0.5, 0.5]^d$, and $\text{loss}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ on each image in (4) follows from [CW17]. Note that $L_2(\theta) = 0$ when all the selected images are successfully attacked. The noisy loss observation $\ell(\theta, \omega)$ in (1) is

$$\ell(\theta, \omega) = \kappa||\theta||_2^2 + \frac{1}{J}\sum_{j=1}^{J} \text{loss}(\zeta_{i_j(\omega)} + \theta), \tag{5}$$

for $J \leq I$, and the J indexes $\{i_1(\omega), \cdots, i_J(\omega)\}$ are i.i.d. uniformly drawn from $\{1, \cdots, I\}$ (without replacement).

Algo	$\mathbb{E}[L(\hat{\theta}_K)]$	$\{\mathrm{Var}[L(\hat{\theta}_K)]\}^{\frac{1}{2}}$	$\mathbb{E}[L_2(\hat{\theta}_K)]$
AdaMM	185.96	16.88	40.95
HARP	138.22	18	12.50

Table: Performance of ZO-AdaMM and HARP in terms of loss after $K = 1000$ iterations averaged across 25 independent replicates. The loss function $L(\cdot)$ is the sum of the magnitude cost $L_1(\cdot)$ and the attack loss $L_2(\cdot)$. Here $L_2(\cdot)$ measures the attack loss on $I = 100$ images of the letter one, and its *noisy* query is evaluated using a batch-size of one. A close-to-zero $L_2(\cdot)$ loss is equivalent to a close-to-one attack success rate.

Algo	$\mathbb{E}[L(\hat{\theta}_K)]$	$\{\mathrm{Var}[L(\hat{\theta}_K)]\}^{\frac{1}{2}}$	$\mathbb{E}[L_2(\hat{\theta}_K)]$
AdaMM	56.95	6.89	11.75
HARP	18.46	1.37	0.13

Table: Here $L_2(\cdot)$ measures the attack loss on $I = 10$ images of the letter three, and its ZO query is noise-free.

Summary

We propose HARP that feeds second-order approximation to $\mathbf{\Sigma}_k$, less sensitive to ill-conditioning.

- framework allowing $\mathcal{F}_k$-measurable perturbation covariance
- second-order info not only gets into parameter update but also *search scaling/direction* in HARP
- asym. rate of convergence remains the same, yet RMS is smaller

Future Work

- this framework can be generalized to scenario where 1st-order oracles are available—Hessian can be estimated using three noisy gradients
- other user-specified structure for $\mathbf{\Sigma}_k$

Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh.
Query-efficient hard-label black-box attack: An optimization-based approach.
In *International Conference on Learning Representations*, 2018.

Nicholas Carlini and David Wagner.
Towards evaluating the robustness of neural networks.
In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh.
Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models.
In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

Yu M Ermol'ev.
On the method of generalized stochastic gradients and quasi-fejér sequences.
Cybernetics, 5(2):208–220, 1969.

Yuri Ermoliev.
Stochastic quasigradient methods and their application to system optimization.
Stochastics: An International Journal of Probability and Stochastic Processes, 9(1-2):1–36, 1983.

V Ya Katkovnik and KULCHITS. OY.
Convergence of a class of random search algorithms.
Automation and Remote Control, 33(8):1321–1326, 1972.

James C. Spall.
Multivariate stochastic approximation using a simultaneous perturbation gradient approximation.
IEEE transactions on automatic control, 37(3):332–341, 1992.

James C Spall.
Adaptive stochastic approximation by the simultaneous perturbation method.
IEEE transactions on automatic control, 45(10):1839–1853, 2000.

James C Spall.
Feedback and weighting mechanisms for improving jacobian estimates in the adaptive simultaneous perturbation algorithm.
IEEE Transactions on Automatic Control, 54(6):1216–1229, 2009.

Jingyi Zhu.
Hessian estimation via stein's identity in black-box problems.
In *2nd Conference on Mathematical and Scientific Machine Learning*, 2021.

J. Zhu, L. Wang, and J. C. Spall.
Efficient implementation of second-order stochastic approximation algorithms in high-dimensional problems.
IEEE Transactions on Neural Networks and Learning Systems, in press at http://dx.doi.org/10.1109/TNNLS.2019.2935455, 2019.