Hessian Estimation via Stein's Identity in Black-Box Problems

Jingyi Zhu

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

August 16, 2021

Minimization Using *Few* Zeroth-Order Queries

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

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

Stochastic Approximation (SA) Algorithms

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

$$\text{2nd-order}: \hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{H}}_k^{-1} \hat{\boldsymbol{g}}_k(\hat{\theta}_k). \tag{3}$$

w/ a_k is stepsize, both $\hat{\boldsymbol{g}}_k$ and $\hat{\boldsymbol{H}}_k$ are approximation using ZO queries.

Comparison of 1st & 2nd methods

Localized model of $f(\theta)$ within $\{\theta + \boldsymbol{d} : ||\boldsymbol{d}|| \leq \delta\}$

$$f(\theta) + \boldsymbol{d}^T \boldsymbol{g}(\theta) + \boldsymbol{d}^T \boldsymbol{B}(\theta)\boldsymbol{d}/2 \tag{4}$$

Letting curvature matrix $\boldsymbol{B}(\theta) = \mathscr{L}_2 \boldsymbol{I}$ motivates (2) and $\boldsymbol{B}(\theta) = \boldsymbol{H}(\theta)$ motivates (3) where $\mathscr{L} \geq \sup_{||\boldsymbol{d}|| \leq \delta} ||\boldsymbol{H}(\theta + \boldsymbol{d})||$.

1. model-trust radius: Levenberg-Marquardt damping technique for (3)
2. computing cost: affordable *storage*, *computation*, and *inversion*.

Benefits of (3) over (2)

- offers faster convergence when $\hat{\theta}_k$ is near θ^*
- eliminates the need for tuning *some* hyperparameters
- local curvature exploitation (preconditioning)
- parameter remains intact under linear mapping

Hessian estimator using ZO oracles

[Fab71] requires $O(d^2)$ ZO queries per iteration.
[Spa00] 2SPSA costs *four* ZO queries.
[PBFM16] 2RDSA costs *three* ZO queries, but with contrived constants.

- [MG15, WMGL17, ABC$^+$19] use first-order oracles
- [ABH17] uses Hessian-vector-product oracle
- [SDPG14, BHNS16, SS19] use second-order oracle

Core Budget Indicator

ZO query complexity to achieve certain level of accuracy.
Besides, floating point operations per iteration may be important.

Stein's Identity

Assume random vector $\boldsymbol{X}$ has density function $p(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$. Under certain conditions

$$\mathbb{E}\left\{q(\boldsymbol{X})[p(\boldsymbol{X})]^{-1}\nabla p(\boldsymbol{X})\right\} = -\mathbb{E}[\nabla q(\boldsymbol{X})]\,, \tag{5}$$

$$\mathbb{E}\left\{q(\boldsymbol{X})[p(\boldsymbol{X})]^{-1}\nabla^2 p(\boldsymbol{X})\right\} = \mathbb{E}[\nabla^2 q(\boldsymbol{X})]\,. \tag{6}$$

Other forms exist for discrete distributed $\boldsymbol{X}$.

$$\hat{\boldsymbol{H}}_k = \begin{cases} c_k^{-2} F_k^{+}(\boldsymbol{u}_k\boldsymbol{u}_k^T - \boldsymbol{I})\,, & \text{(7a)} \\ c_k^{-2}(F_k^{+} - F_k)(\boldsymbol{u}_k\boldsymbol{u}_k^T - \boldsymbol{I})\,, & \text{(7b)} \\ (2c_k^2)^{-1}(F_k^{+} + F_k^{-})(\boldsymbol{u}_k\boldsymbol{u}_k^T - \boldsymbol{I})\,, & \text{(7c)} \\ (2c_k^2)^{-1}(F_k^{+} + F_k^{-} - 2F_k)(\boldsymbol{u}_k\boldsymbol{u}_k^T - \boldsymbol{I})\,. & \text{(7d)} \end{cases}$$

$F_k^{\pm} \equiv F(\hat{\theta}_k \pm c_k\boldsymbol{u}_k, \omega_k^{\pm})$, $\boldsymbol{u}_k$ follows multivariate standard normal distribution, c_k is differencing magnitude.

- On the basis of the same convergence rate, our estimator requires *three* queries, while 2SPSA needs *four*. Besides, we require generating *one* perturbation vector and tuning *one* differencing magnitude, while 2SPSA needs *two*.
- Our estimator is naturally symmetric, while 2SPSA requires manual symmetrization.
- We require "thrice cont' differentiable w/ Lipschitz continuous 3rd-order derivatives", while 2SPSA requires "four-times continuously differentiable w/ bounded fourth-order derivatives".
- Thanks to Stein's identity, proof is simplified.
  Following proofs of 2SPSA will not give fastest convergence.
- The smoothing scheme for estimating the Hessian estimator is generalized.

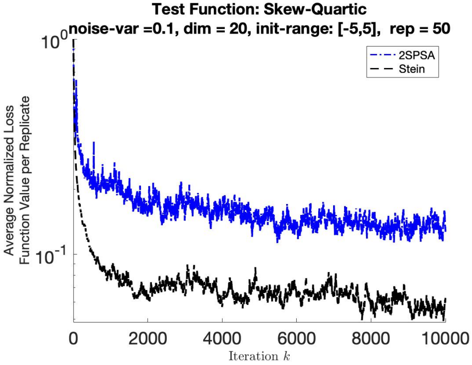

Figure: Performance of ours and 2SPSA in terms of normalized distance $[f(\hat{\theta}_k)-f(\theta^*)]/[f(\hat{\theta}_0)-f(\theta^*)]$ average across 50 independent replicates. Both algorithms use twelve ZO queries per iteration, so query complexity aligns with iteration complexity. The underlying loss function is the skew-quartic function with $d = 20$, and the noisy observation is corrupted by $N(0, 0.1)$ noise.

[HTC20] uses PHISHING dataset for black-box classification.

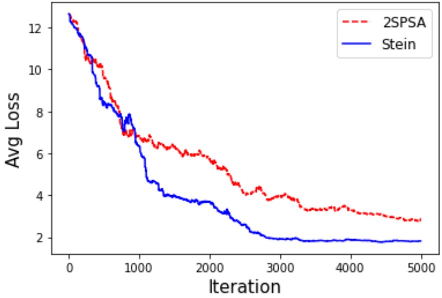

Figure: Performance of ours and 2SPSA in terms of the true loss function average across 25 independent replicates. Both algorithms use twelve ZO queries per iteration, so query complexity aligns with iteration complexity. A zero loss function is equivalent to 100% classification correctness.

Summary

Stein's Identity helps in validating Hessian estimators

- reduced ZO query compared with 2SPSA
- reduced random perturbation generation, gain tuning

Future Work

- extension to case where unbiased direct measurements of gradient information is available
- extension for other possible distribution for random perturbation

Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, and Yi Zhang.
Efficient full-matrix adaptive regularization.
In *International Conference on Machine Learning*, pages 102–110, 2019.

Naman Agarwal, Brian Bullins, and Elad Hazan.
Second-order stochastic optimization for machine learning in linear time.
The Journal of Machine Learning Research, 18(1):4148–4187, 2017.

Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer.
A stochastic quasi-Newton method for large-scale optimization.
SIAM Journal on Optimization, 26(2):1008–1031, 2016.

V Fabian.
Stochastic approximation, optimization methods in statistics.
In J. S. and Rustigi, editor, *Optimizing Methods in Statistics*, pages 439–470. Academic Press, New York, 1971.

Feihu Huang, Lue Tao, and Songcan Chen.
Accelerated stochastic gradient-free and projection-free methods.
In *International Conference on Machine Learning*, pages 4519–4530. PMLR, 2020.

James Martens and Roger Grosse.
Optimizing neural networks with kronecker-factored approximate curvature.
In *International conference on machine learning*, pages 2408–2417, 2015.

LA Prashanth, Shalabh Bhatnagar, Michael Fu, and Steve Marcus.
Adaptive system optimization using random directions stochastic approximation.
IEEE Transactions on Automatic Control, 62(5):2223–2238, 2016.

Jascha Sohl-Dickstein, Ben Poole, and Surya Ganguli.
Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods.
In *Proceedings of the International Conference on Machine Learning*, pages 604–612, Beijing, China, 21–26 June 2014.

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

Samer S Saab and Dong Shen.
Multidimensional gains for stochastic approximation.
IEEE Transactions on Neural Networks and Learning Systems, 31(5):1602–1615, 2019.

Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu.
Stochastic quasi-newton methods for nonconvex stochastic optimization.
SIAM Journal on Optimization, 27(2):927–956, 2017.