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

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

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

Stochastic Approximation (SA) Algorithms

1st-order: \( \hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k) \).  

\[
\hat{g}_k(\hat{\theta}_k) = \frac{\ell(\hat{\theta}_k + c_k \Delta_k, \omega_k^+) - \ell(\hat{\theta}_k - c_k \Delta_k, \omega_k^-)}{2c_k} m_k(\Delta_k). \]

w/ \( c_k \) is differencing magnitude, \( \Delta_k \) is perturbation, \( m_k(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d \).
RDSA (random direction) [Erm69, Erm83]

\[ \Delta_k \sim \text{Unif}(\mathbb{S}) \quad \text{and} \quad m_k(\Delta_k) = d\Delta_k \]

SPSA (simultaneous perturbation) [Spa92]

\[ \Delta_k \sim [\text{Unif}\{-1, 1\}]^d, \quad m_k(\Delta_k) = \Delta_k \]

SFSA (smoothed functional) [KO72]

\[ \Delta_k \sim \mathcal{N}(0, I) \quad \text{and} \quad m_k(\Delta_k) = \Delta_k \]

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

\[ \Delta_k \sim \text{Unif}\{e_1, \cdots, e_d\} \quad \text{and} \quad m_k(\Delta_k) = d\Delta_k \]

What if \( \Delta_k \) has zero-mean and \( \Sigma_k^{-1} \)-covariance for \( \Sigma_k \succ 0 \)?

- Data-driven \( \Sigma_k \) can be \( \mathcal{F}_k \)-measurable random variable
- User-specified \( \Sigma_k \) can use domain knowledge
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. $E_{\Delta}[\hat{g}(\hat{\theta})]$ equals true gradient $g(\hat{\theta}) = [100, 1]^T$. But the variance is in the order of $10^4$ assuming noise-free ZO queries.

- HARP draws $\Delta$ from 0-mean and $\Sigma^{-1}$-covariance distribution with $\Sigma = \hat{H}(\hat{\theta})$. Still unbiased, but covariance matrix norm is $2 \times 10^2$.

$\Sigma = I$ has zero off-diagonal elements implies that each component of $\Delta$ 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 $4 \times 10^4$ and that of HARP is $8 \times 10^2$. 
HARP handles scaling & correlation

\( \Delta_k \) follows a dist. w/ mean 0 and cov. \( \hat{H}_k^{-1} \), and \( m_k(\Delta_k) = \hat{H}_k \Delta_k \).

Theoretical Guarantee

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

- Expectation of estimation error goes to zero at the same rate.
- Variance gets smaller while using \( \Sigma_k = \hat{H}_k \).

\(^a k^{-1/3}\) when \( \omega_k^+ \) and \( \omega_k^- \) are independent and identically distributed (IID).
\(^b k^{-1/2}\) 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 $\Sigma_k$

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

$$
\begin{align*}
\min_\theta L(\theta) &= \kappa \|\theta\|_2^2 + \frac{1}{I} \sum_{i=1}^{I} \text{loss}(\zeta_i + \theta), \\
&\equiv L_1(\theta) \\
&\equiv L_2(\theta)
\end{align*}
$$

where the constraint is to normalize the resulting pixels within $[-0.5, 0.5]^d$, and $\text{loss}() : \mathbb{R}^d \mapsto \mathbb{R}$ 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

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

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).
Numerical Study  Image Attack

<table>
<thead>
<tr>
<th>Algo</th>
<th>$\mathbb{E}[L(\hat{\theta}_K)]$</th>
<th>${\text{Var}[L(\hat{\theta}_K)]}^{\frac{1}{2}}$</th>
<th>$\mathbb{E}[L_2(\hat{\theta}_K)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAMM</td>
<td>185.96</td>
<td>16.88</td>
<td>40.95</td>
</tr>
<tr>
<td>HARP</td>
<td>138.22</td>
<td>18</td>
<td>12.50</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>Algo</th>
<th>$\mathbb{E}[L(\hat{\theta}_K)]$</th>
<th>${\text{Var}[L(\hat{\theta}_K)]}^{\frac{1}{2}}$</th>
<th>$\mathbb{E}[L_2(\hat{\theta}_K)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAMM</td>
<td>56.95</td>
<td>6.89</td>
<td>11.75</td>
</tr>
<tr>
<td>HARP</td>
<td>18.46</td>
<td>1.37</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**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 $\Sigma_k$, 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 $\Sigma_k$
Query-efficient hard-label black-box attack: An optimization-based approach.

Nicholas Carlini and David Wagner.
Towards evaluating the robustness of neural networks.

Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models.

Yu M Ermol’ev.
On the method of generalized stochastic gradients and quasi-fejér sequences.

Yuri Ermoliev.
Stochastic quasigradient methods and their application to system optimization.

V Ya Katkovnik and KULCHITS. OY.
Convergence of a class of random search algorithms.

James C. Spall.
Multivariate stochastic approximation using a simultaneous perturbation gradient approximation.

James C Spall.
Adaptive stochastic approximation by the simultaneous perturbation method.
James C Spall.
Feedback and weighting mechanisms for improving jacobian estimates in the adaptive simultaneous perturbation algorithm.

Jingyi Zhu.
Hessian estimation via stein’s identity in black-box problems.

J. Zhu, L. Wang, and J. C. Spall.
Efficient implementation of second-order stochastic approximation algorithms in high-dimensional problems.