Adaptive Step Sizes (for Nystrom and rSVD)

Fast and Randomized Principle Component Analysis

Theorem 1. Define the $d \times d$ matrix $C$ as $\frac{1}{n}X^T X - \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$, and let $V_k$ denote the $d \times k$ matrix composed of the eigenvectors corresponding to the largest $k$ eigenvalues. Suppose that

- $\text{max} \| x_i \|^2 \leq r$ for some $r > 0$,
- $C$ has eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d$, where $\lambda_d = \lambda_1 = \lambda$ for some $\lambda > 0$, and
- $k = \| V_k \|_{F} \leq \frac{1}{\lambda}$

Let $\delta, \epsilon \in (0, 1)$ be fixed. If we run the algorithm with any epoch length parameter $m$ and step size $\eta$, such that

$$\eta \leq \frac{c\delta}{\lambda \epsilon}, \quad m \geq \frac{c\log(2/\delta)}{\eta^2}, \quad km\epsilon^2 + r \sqrt{m} \lambda \log(2/\delta) \leq \epsilon^2$$

(9)

(where $c, c', c''$ designate certain positive numerical constants), and for $T = \frac{\log(1/\delta)}{\log\log(1/\delta)}$ epochs, then with probability at least $1 - 1 \cdot \log(1/\delta)/9$, it holds that

$$k = \| V_k \|_{F} \leq \epsilon.$$

On Convergence of Stochastic Gradient Descent with Adaptive Step Sizes, from Li and Orabona ’19

Theorem 4: Assume (H1, H3, H4'). Let $\eta_t$ be our global generalized AdaGrad stepsize before, where $\alpha, \beta > 0$ and $c \in (0, 1/2)$, and $4\alpha M < \beta^{1/2 + \epsilon}$. Then the iterates of SGD satisfy the following bound:

$$\mathbb{E} \left[ \min_{1 \leq t \leq T} \left\| \nabla f(x_t) \right\|^{1 - 2\epsilon} \right] \leq \frac{1}{1 - 2\epsilon} \alpha \left( 2^{1/2 + \epsilon} \gamma, 2^{1/2 + \epsilon} (\beta + 2 T \sigma^2)^{1/4 - \epsilon} \gamma^{1/2 - \epsilon} \right).$$

Discussion

4.2 Gaussian Process Regression

Similarly, we also apply different methods to Gaussian process regression. We randomly generate 200 training points, and construct the Gaussian kernel surrogate of it by calculating the function $f(x) = \sum_{i=1}^{n} \frac{c}{\sqrt{2\pi}} e^{-\frac{(x_i - x)^2}{2\sigma^2}}$.

Empirical Results Suggest Quasi-Monte Carlo Sampling Increases Accuracy in OpenOA AEP

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meter</td>
<td>$N(1, 0.005)$</td>
</tr>
<tr>
<td>Loss</td>
<td>$N(1, 0.05)$</td>
</tr>
<tr>
<td>Windiness</td>
<td>$\text{Uniform}(10, 21)$</td>
</tr>
<tr>
<td>Loss Threshold</td>
<td>$\text{Uniform}(10, 21)$</td>
</tr>
<tr>
<td>Reanalysis Product</td>
<td>$\text{Uniform}(0, 1)$</td>
</tr>
<tr>
<td>Compute Time</td>
<td></td>
</tr>
</tbody>
</table>

A Nonlinear Extension to Kalman GD:

Unsuccessfully Attempting to Combine Uncertainty Quantification with Variance Reduction

$$X = \begin{bmatrix} x(t_1) & x(t_2) & x(t_3) & \ldots & x(t_n) \end{bmatrix}$$

Accelerated Proper Orthogonal Decomposition for Turbulent Flows

(Aviral Prakash)

So why is filtering stochastic gradients a bad idea?

Ideal Kalman SG

$$\begin{bmatrix} \hat{\theta} \\ \nabla f(\theta) \end{bmatrix} = \begin{bmatrix} 0 & -\alpha \nabla^2 f(\theta) \end{bmatrix} \begin{bmatrix} \hat{\theta} \\ \nabla f(\theta) \end{bmatrix}$$

Observation Model

$$y = Hx = [0 \ I]x + \eta$$

Actual Kalman SG

$$\Delta \hat{\theta} = \begin{bmatrix} 0 & -\alpha \nabla^2 f(\theta) \end{bmatrix} \begin{bmatrix} \Delta \hat{\theta} \\ \Delta \nabla f(\theta) \end{bmatrix}$$

Observation Model

$$y_k = Hx_k = [0 \ I]x_k' + \eta$$

APP5650

“Randomized Algorithms” Prof. Becker, fall 2021

Student projects

Student backgrounds:

- Applied Math (MS, PhD)
- Math (PhD)
- Computer Science (MS, PhD)
- Aerospace Engineering (PhD)