Unbiased Stochastic Gradient Descent for Strongly Correlated Time Series Data

Tim Anthony

Supervisor: Xin Tong
National University of Singapore

July 1, 2024

Outline

Formal optimization problem

- Quantity of interest: $y_t = y(w^*, x_t, \xi_t)$
- Parameterized prediction of y_t: $\hat{y}_t = \hat{y}(w, x_t)$
- Expected loss function to quantify the error: $F(w) = E_{X,Y}[f(w, X, Y)]$
- Optimization objective: find $w^* = \underset{w}{argmin}\, F(w)$

Vanilla SGD

- Costly using gradient descent: $w_{t+1} = w_t - \eta_{t+1} \nabla_w F(w_t)$
- Idea: Use sample estimation of $\nabla_w F(w_t)$ as $\nabla_w f(w_t, x_{t+1}, y_{t+1})$
  $\rightarrow$ SGD: $w_{t+1} = w_t - \eta_{t+1} \nabla_w f(w_t, x_{t+1}, y_{t+1})$
- Key assumption: (X_t, Y_t) are i.i.d, hence by LLN

$$E_t[\nabla_w f(w, x_{t+1}, y_{t+1})] = \nabla_w F(w)$$

Problem with correlated data

- We consider the setting when the x-data is generated by a stationary AR(1) process
- Normal for financial data such as portfolio returns
- Hence X_t-data are auto-correlated, and (X_t, Y_t) are no longer i.i.d.
- Sample gradient biased: $E_t[\nabla_w f(w, x_{t+1}, y_{t+1})] \neq \nabla_w F(w)$
- For highly correlated data, standard SGD will suffer significantly

Main idea - Unbiased SGD

- Find Q_t such that:

$$Q_t E_t[\nabla_w f(w, x_{t+1}, y_{t+1})] = \nabla_w F(w)$$

- Perform the following unbiased SGD updates:

$$w_{t+1} = w_t - \eta_{t+1} Q_t \nabla_w f(w_t, x_{t+1}, y_{t+1})$$

Linear Univariate Case - Model

Data generating process:

- x-data: $x_{t+1} = \rho x_t + \epsilon_{t+1}, \quad \rho < 1$ *AR(1)*
- y-data: $y_{t+1} = w^* x_{t+1} + \xi_{t+1}$ *Linear regression model*

Prediction:

- $\hat{y}_{t+1} = w x_{t+1}$ *Linear regression estimate*

Loss function:

- $f(w, x, y) = \frac{1}{2}(wx - y)^2$ *L_2-loss*

Linear Univariate Case - Finding Q_t

- Want to find Q_t such that: $Q_t E_t[\nabla_w f(w, X_{t+1}, Y_{t+1})] = E[\nabla_w f(w, X_{t+1}, Y_{t+1})]$

- Can calculate:
  - $E_t[\nabla_w f(w, X_{t+1}, Y_{t+1})] = [\rho^2 x_t^2 + \sigma_\epsilon^2](w - w^*)$
  - $E[\nabla_w f(w, X_{t+1}, Y_{t+1})] = \sigma_x^2(w - w^*)$

- Q_t given as:

$$Q_t = \frac{\sigma_x^2}{\rho^2 x_t^2 + \sigma_\epsilon^2}$$

Linear Univariate Case - Intuition of Q_t

For stationary AR(1) process, $\sigma_\epsilon^2 = (1-\rho^2)\sigma_x^2$:

$$Q_t = \frac{\sigma_x^2}{\sigma_x^2 + \rho^2(x_t^2 - \sigma_x^2)}$$

Hence,

$$Q_t = \begin{cases} >1, & \textit{if } x_t^2 < \sigma_x^2 \\ =1, & \textit{if } x_t^2 = \sigma_x^2 \\ <1, & \textit{if } x_t^2 > \sigma_x^2 \end{cases}$$

While the conditionally expected error is given as

$$E_t[error_{t+1}] = E_t[\hat{y}_{t+1} - y_{t+1}] = \rho x_t(w_t - w^*)$$

Linear Univariate Case - Results

Convergence to $E[(w_{t+1} - w^*)^2] \leq \Delta$:

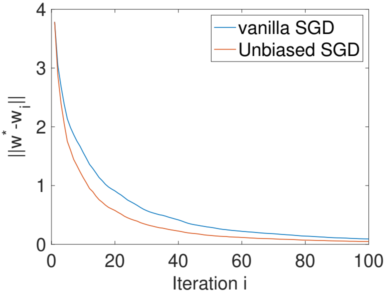
Vanilla SGD:

- $t \geq \Omega\left(\frac{log(\Delta)}{\Delta(1-\rho^2)^2}\right)$

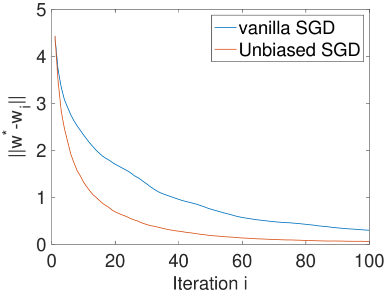
Unbiased SGD:

- $t \geq \Omega\left(\frac{log(\Delta)}{\Delta(1-\rho^2)}\right)$

Numerical:



(a) $\rho = 0.8$



(b) $\rho = 0.95$

Linear Multivariate Case - Model

Data generating process:

- x-data: $x_{t+1} = Px_t + \epsilon_{t+1}, \quad \lambda_{max}(P) < 1$ *AR(1)*
- y-data: $y_{t+1} = x_{t+1}^T w^* + \xi_{t+1}$ *Linear regression model*

Prediction:

- $\hat{y}_{t+1} = x_{t+1}^T w$ *Linear regression estimate*

Loss function:

- $f(w, x, y) = \frac{1}{2}(x^T w - y)^2$ *L_2-loss*

Linear Multivariate Case - Finding Q_t

- Want to find Q_t such that: $Q_t E_t[\nabla_w f(w, X_{t+1}, Y_{t+1})] = E[\nabla_w f(w, X_{t+1}, Y_{t+1})]$
- Can calculate:
  - $E_t[\nabla_w f(w, X_{t+1}, Y_{t+1})] = [P x_t x_t^T P^T + \Sigma_\epsilon](w - w^*)$
  - $E[\nabla_w f(w, X_{t+1}, Y_{t+1})] = \Sigma_x (w - w^*)$
- Q_t given as:

$$Q_t = \Sigma_x [P x_t x_t^T P^T + \Sigma_\epsilon]^{-1}$$

Linear Multivariate Case - Results

Convergence to $E[||w_{t+1} - w^*||^2] \leq \Delta$:

Vanilla SGD:

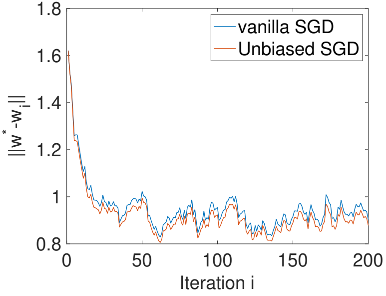
- $t \geq \Omega\left(\frac{log(\Delta)}{\Delta(\lambda_\epsilon^{min})^2}\right)$

Unbiased SGD:

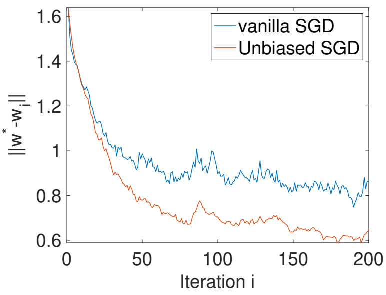
- $t \geq \Omega\left(\frac{log(\Delta)}{\Delta\lambda_\epsilon^{min}}\right)$

Note: $\Sigma_\epsilon = \Sigma_x - P\Sigma_x P^T \to 0$ as $\lambda_P^{min} \to 1$

Numerical:



(a) $\lambda_P^{min} = 0.1$



(b) $\lambda_P^{min} = 0.9$

General Unbiased SGD - Generalizing further

- Want to consider more general class of functions f
- Conditions:
  - $\nabla_w f(w, x, y)$ and $H_f(w, x, y)$ known
  - $E_t[\nabla_w f(w^*, X_{t+1}, Y_{t+1})] = E[\nabla f(w^*, X_{t+1}, Y_{t+1})] = 0$

General Unbiased SGD - Finding Q_t

How to find Q_t such that:

$$Q_t E_t[\nabla_w f(w_t, X_{t+1}, Y_{t+1})] = E[\nabla_w f(w, X_{t+1}, Y_{t+1})]$$

Main idea:
If

$$\begin{cases} Q_t E_t[\nabla_w f(w_t, X_{t+1}, Y_{t+1})] \approx Q_t A c \\ E[\nabla_w f(w, X_{t+1}, Y_{t+1})] \approx B c \end{cases}$$

then

$$Q_t \approx A^{-1} B.$$

General Unbiased SGD - Calculating Q_t

Taylor expansion around w^*:

$$\nabla_w f(w, X_{t+1}, Y_{t+1}) \approx \nabla_w f(w^*, X_{t+1}, Y_{t+1}) + H_{f_w}(w^*, X_{t+1}, Y_{t+1})(w^* - w).$$

By

$$E_t[\nabla_w f(w^*, X_{t+1}, Y_{t+1})] = E[\nabla_w f(w^*, X_{t+1}, Y_{t+1})] = 0$$

we can identify

$$\begin{cases} A = E_t[H_{f_w}(w^*, X_{t+1}, Y_{t+1})] \\ B = E[H_{f_w}(w^*, X_{t+1}, Y_{t+1})] \\ c = (w^* - w) \end{cases}$$

and hence

$$Q_t \approx E_t[H_{f_w}(w^*, X_{t+1}, Y_{t+1})]^{-1} E[H_{f_w}(w^*, X_{t+1}, Y_{t+1})]$$

Thank you