Variance Reduced Training with Stratified Sampling for Forecasting Models (Supplementary Materials)

A. Details in Experiments and Addtional Results

A.1. Practical Version of SCott

We first discuss in details on the pratical version of SCott as mentioned in Section 6. The full description is shown in Algorithm 2. The main difference between the plain SCott and Algorithm 2 is the latter treats the period K of performing stratified as a constant, while adaptively altering K based on a stopping criteria γ. Adopting such technique, despite being complicated in theory, allows us to adaptively perform the stratified sampling based on the progress of the training. The hyperparameter γ can then be obtained via standard tuning algorithms such as grid search and random search.

Algorithm 2 The practical version of SCott, where we apply the early stopping technique, instead of choosing K based on Geometric distribution.

Require: Total number of iterations T, learning rate $\{\alpha_t\}_{1\leq t\leq T}$, initialize $\boldsymbol{\theta}^{(0,0)}$, strata: $\{\mathcal{D}_i\}_{1\leq i\leq B}$, initialized selection of K, hyperparameter γ.

for $t = 0, 1, \cdots, T-1$ **do**
  Sample a $\xi_i^{(t)}$ from stratum i and perform stratified sampling (with $w_i = |\mathcal{D}_i|/|\mathcal{D}|$): $\boldsymbol{g}^{(t,0)} = \sum_{i=1}^{B} w_i \nabla f_{\xi_i^{(t)}}(\boldsymbol{\theta}^{(t,0)})$.
  for $k = 0, 1, \cdots, K-1$ **do**
    Sample $\xi^{(t,k)}$ from $\mathcal{D}$.
    Compute the update $\boldsymbol{v}^{(t,k)}$ as $\nabla f_{\xi^{(t,k)}}(\boldsymbol{\theta}^{(t,k)}) - \nabla f_{\xi^{(t,k)}}(\boldsymbol{\theta}^{(t,0)}) + \boldsymbol{g}^{(t,0)}$.
    Update the parameters as $\boldsymbol{\theta}^{(t,k+1)} = \boldsymbol{\theta}^{(t,k)} - \alpha_t \boldsymbol{v}^{(t,k)}$.
    if $\|\boldsymbol{v}^{(t,k)}\|^2 \leq \gamma \|\boldsymbol{v}^{(t,0)}\|^2$ **then**
      break
    end if
  end for
  Set $\boldsymbol{\theta}^{(t+1,0)} = \boldsymbol{\theta}^{(t,k+1)}$.
end for
return Sample $\hat{\boldsymbol{\theta}}^{(T)}$ from $\{\boldsymbol{\theta}^{(t,0)}\}_{t=0}^{T-1}$ with $\mathbb{P}(\hat{\boldsymbol{\theta}}^{(T)} = \boldsymbol{\theta}^{(t,0)}) \propto \alpha_t B$

A.2. Synthetic Dataset Generation

In this subsection, we introduce the details of generating synthetic time series dataset as used in the experiments. We first set the context length to be 72 and prediction length to be 24. This synthetic dataset contains 4 time series with different patterns on their time horizon. We start from the definition of four types of patterns: $\boldsymbol{t} = [1, 2, \cdots, 23, 24] \in \mathbb{R}^{24}$, $P_1 = \sin(\boldsymbol{t}) \in \mathbb{R}^{24}$, $P_2 = \boldsymbol{t} \in \mathbb{R}^{24}$, $P_3 = \boldsymbol{t}^2 \in \mathbb{R}^{24}$, $P_4 = \sqrt{\boldsymbol{t}} \in \mathbb{R}^{24}$, where all the transformations are element-wise, e.g. $\sin(\boldsymbol{t}) = [\sin(1), \sin(2), \cdots, \sin(23), \sin(24)]$. And the four patterns are four different time series slices of length 24 that maps $\boldsymbol{t}$ to different values via transformations of *sin*, *linear*, *quadratic*, *square root*, respectively. With these patterns, the time series data are then constructed via concatenating the patterns with different orders on the time horizon. Specifically,

$$TS_1 = \underbrace{[P_1, P_2, P_3, P_4, \cdots, P_1, P_2, P_3, P_4]}_{[P_1,P_2,P_3,P_4] \text{ repeats 2K times}} + \mathcal{N}(0, \mathbf{1}), \quad \text{where} \quad \mathcal{N}(0, \mathbf{1}) \in \mathbb{R}^{192K}$$

$$TS_2 = \underbrace{[P_4, P_3, P_2, P_1, \cdots, P_4, P_3, P_2, P_1]}_{[P_4,P_3,P_2,P_1] \text{ repeats 2K times}} + \mathcal{N}(0, \mathbf{1}), \quad \text{where} \quad \mathcal{N}(0, \mathbf{1}) \in \mathbb{R}^{192K}$$

$$TS_3 = \underbrace{[P_1, P_3, P_2, P_4, \cdots, P_1, P_3, P_2, P_4]}_{[P_1,P_3,P_2,P_4] \text{ repeats 2K times}} + \mathcal{N}(0, \mathbf{1}), \quad \text{where} \quad \mathcal{N}(0, \mathbf{1}) \in \mathbb{R}^{192K}$$

$$TS_4 = \underbrace{[P_4, P_2, P_3, P_1, \cdots, P_4, P_2, P_3, P_1]}_{[P_4,P_2,P_3,P_1] \text{ repeats 2K times}} + \mathcal{N}(0, \mathbf{1}), \quad \text{where} \quad \mathcal{N}(0, \mathbf{1}) \in \mathbb{R}^{192K}$$

where $\mathcal{N}(0, \mathbf{1})$ denotes a random vector where each coordinate is sampled from a normal distribution. We can see each time series is repeating a distinct order of the four patterns, forming a temporal pattern, on its time horizon. We let such temporal pattern repeat 2K times on the time horizon of each time series. Finally, a Gaussian noise is added to each time series to capture randomness. After generating the time series, we then extract the training examples via a sliding window which slides 24 timestamps[5] between adjacent training examples. With a simple calculation, the total number of training examples in this dataset is 32K.

Stratification. From the data generation process, we can see in each time series contains exactly four types of mapping: take TS_1, the type of mappings on its time horizon are $\forall i = 1, 2, 3, 4$,

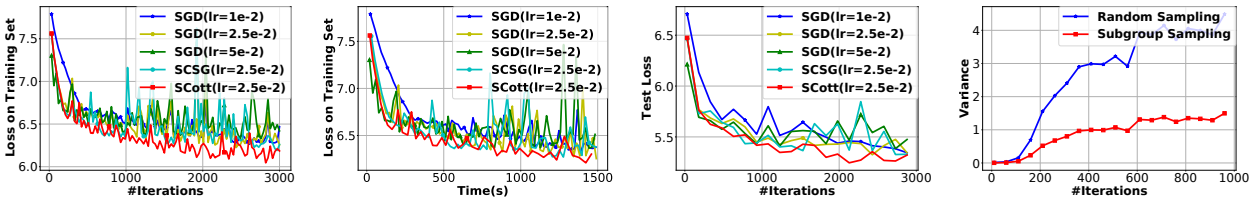
$$\text{Mapping i: } [P_{(i+1) \bmod 4}, P_{(i+2) \bmod 4}, P_{(i+3) \bmod 4}] \to P_i$$

Same conclusions can be drawn on other time series. And then we can stratify all the training examples via a simple judgemental policy: the training examples that belonging to the same time series and having same pattern in their prediction range are clustered to the same stratum. The total number of strata is then 16.

A.3. Additional Results

A.3.1. Results additional to the main paper settings

In the main paper, we show the convergence curves of training MLP on Traffic dataset and training NBEATS on Electricity dataset. Here, we provide the additional curves of training MLP on Electricity dataset.



(a) Training Loss / # Iterations MLP - Electricity Dataset

(b) Training Loss / Time(s) MLP - Electricity Dataset

(c) Test Loss / # Iterations MLP - Electricity Dataset

(d) Variance of Samplers MLP - Electricity Dataset

Figure 4. Additional Results of MLP on Electricity Dataset.

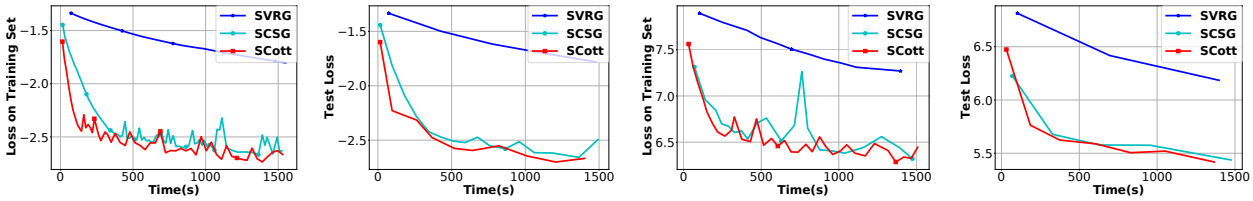
A.3.2. Applying Early Stopping Technique to SCSG and SVRG

In this subsection, we investigate how the baseline SCSG and SVRG would perform when the early stopping technique introduced in Section A.1 is applied on them. We rerun the MLP model on Traffic and Electricity dataset, and the fine tuned results are shown in Figure 5. In the literature, SVRG is shown to perform bad on deep learning tasks (Defazio & Bottou, 2018).

A.4. Hyperparameter Tuning

We apply the grid search to tune the hyperparameters in each experiment, the grids for the step sizes are: {0.1, 0.05, 0.025, 0.01, 0.005, 0.0025, 0.001, 0.0005, 0.00025, 0.0001}. We set the weight decay to be 1e-5. For experiments in Section 7.1 and 7.2, the optimal step size is specified. For SCott, we additionally tune γ from $\{0.1, 0.125, 0.15, 0.2\}$, and the optimal

[5] We do this to guarantee each training example can include a completely different pattern in its prediction range. In practice, this can mean extracting training examples by day on a hourly measured time series.



(a) Training Loss w.r.t. Time(s) Traffic Dataset (b) Test Loss w.r.t. Time(s) Traffic Dataset (c) Training Loss w.r.t. Time(s) Electricity Dataset (d) Test Loss w.r.t. Time(s) Electricity Dataset

Figure 5. Applying the early stopping technique on SCSG and SVRG.

choice of γ on synthetic dataset is 0.125. Hyperparameters in other settings is shown in Table A.4. For the experiment in section 7.3, we have two extra hyperparameters β_1 and β_2, we set the grids to be $\{0.9, 0.99, 0.999\}$. Finally, for γ, we extend the grids range to $\{0.1, 0.125, 0.15, 0.2, 0.4, 0.8\}$. The optimal choice of hyperparameter for each settings are shown in Table A.4.

Table 4. Hyperparameters used for experiments on real-world applications. The format of the hyperparameters is shown as the following: the first value is the optimal choice of step size. For SCott-type optimizers, the last value is the optimal choice for γ. Additionally, the β_1 and β_2 for Adam-type optimizers are set to be 0.9 and 0.999 respectively for optimal performance.

Model	Dataset	Optimizer						
		SGD	SCSG	SCott	Adam	S-Adam	Adagrad	S-Adagrad
MLP	Exchange Rate	5e-3	5e-2	5e-2/0.125	5e-3	5e-3/0.1	2.5e-2	2.5e-2/0.1
	Traffic	2.5e-2	2.5e-2	2.5e-2/0.125	5e-3	5e-3/0.125	5e-3	5e-3/0.125
	Electricity	2.5e-2	2.5e-2	2.5e-2/0.125	5e-3	2.5e-3/0.125	5e-3	2.5e-3/0.125
NBEATS	Exchange Rate	1e-3	1e-3	1e-3/0.1	1e-3	1e-3/0.1	1e-3	1e-3/0.1
	Traffic	1e-4	1e-4	2.5e-3/0.125	1e-4	1e-4/0.125	1e-4	1e-4/0.125
	Electricity	5e-3	1e-2	1e-2/0.125	1e-2	1e-2/0.125	1e-2	1e-2/0.125

B. Technical Proof.

B.1. Heterogeneity Noise with Uniform Sampling

As a supplementary to the main paper, here we investigate another example to illustrate why time series data can be heterogeneous, and how uniform sampling could cause extra noise on such dataset. Consider the simplest AR model with $p = 1$. Without the loss of generality, we assume the parameter is initialized at point 0: $\boldsymbol{\theta}^{(0)} = 0 \in \mathbb{R}$. Now consider a dataset $\mathcal{D}$ contains a single time series that takes the following form:

$$\underbrace{-1}_{t=1}, -\delta, 1, \delta, \underbrace{-1, -\delta, 1, \delta}_{\text{Temporal Pattern}}, \cdots,$$

where $\delta > 0$ denote some constant. We can see in this example a temporal pattern of length 4 is repeating on the time horizon, and the conditional distribution over the timestamp has two types. With some simple analysis, for all the examples whose prediction time t_0 fulfilling $t_0 \bmod 2 = 1$, their global minima are centering around $\theta_1^* = \delta$. We denote all these training examples as $\mathcal{D}_1$. Similarly, for all the t_0 with $t_0 \bmod 2 = 0$, their global minima are centering around $\theta_2^* = -\frac{1}{\delta}$. We denote all these training examples as $\mathcal{D}_2$. In this toy dataset, the heterogeneity comes from the fact that $\mathcal{D}_1$ and $\mathcal{D}_2$ are not gathering around the same global minimia: when δ increases, the distance of two minima $|\delta + 1/\delta|$ centers will increase. We now present the problem with uniform sampling in the following lemma.

Lemma 2. *Consider using uniform sampling on $\mathcal{D}$ to obtain a mini-batch ξ with size of $M = 2$, then when the training examples in ξ are either both sampled from $\mathcal{D}_1$ or both from $\mathcal{D}_2$, with high probability,*

$$Var\left[\nabla f_\xi(\boldsymbol{\theta}^{(0)})\right] = \underbrace{O(\delta^2)}_{\textit{heterogeneity noise}} + o(\delta). \tag{12}$$

On the otherhand, when the training examples in ξ are sampled from both $\mathcal{D}_1$ and $\mathcal{D}_2$ using stratified sampling, with high probability,

$$Var\left[\nabla f_{\xi}(\boldsymbol{\theta}^{(0)})\right] = \underbrace{o(\delta)}_{\textit{model randomness}}. \tag{13}$$

Lemma 2 shows with uniform sampling, the variance of the stochastic gradients is closely related to the samples: If the samples are somewhat similar, the variance on the gradient will suffer from an additional term – to which we informally refer as heterogeneity noise.

Proof to Lemma 2.

Proof. Note that for AR(1), the model only has a single parameter. In this proof, $\boldsymbol{\theta}$ and $[\theta]$ are interchangeable. The loss functions at different time can be expressed as:

$$f_{1,t}(\boldsymbol{\theta}) = f_{1,t}([\theta]) = \begin{cases} (\delta\theta + 1 + \epsilon_t)^2 & t \bmod 4 = 0 \\ (\theta - \delta - \epsilon_t)^2 & t \bmod 4 = 1 \\ (\delta\theta + 1 - \epsilon_t)^2 & t \bmod 4 = 2 \\ (\theta - \delta + \epsilon_t)^2 & t \bmod 4 = 3 \end{cases},$$

Take gradient with respect to the model parameter θ and without the loss of generality, taking $\boldsymbol{\theta}^{(0)} = [0]$, we obtain

$$\nabla f_{1,t}([0]) = \begin{cases} 2\delta + 2\delta\epsilon_t & t \bmod 4 = 0 \\ -2\delta - 2\epsilon_t & t \bmod 4 = 1 \\ 2\delta - 2\delta\epsilon_t & t \bmod 4 = 2 \\ -2\delta + 2\epsilon_t & t \bmod 4 = 3 \end{cases}. \tag{14}$$

As defined by Equation (1), the gradient on the total loss can be expressed as (with out the loss of generality, we set $T = 4\tilde{T}$ where $\tilde{T}$ is an integer.)

$$\nabla f(\boldsymbol{\theta}^{(0)}) = \frac{1}{T}\sum_{t=1}^{T} \nabla f_{1,t}(\boldsymbol{\theta}^{(0)}) = \frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1} - \delta\epsilon_{m+2} + \epsilon_{m+3} + \delta\epsilon_{m+4}). \tag{15}$$

Denote $\mathcal{E}$ as the event of "two training examples in the mini-batch are either both sampled from $\mathcal{D}_1$ or both sampled from $\mathcal{D}_2$". Depend on the event of $\mathcal{E}$, we first obtain when the event $\mathcal{E}$ happens,

$$\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\middle|\mathcal{E}\right] \tag{16}$$

$$= \frac{1}{T^2}\sum_{t,t' \bmod 4=0}\left|2\delta + \delta\epsilon_t + \delta\epsilon_{t'} - \frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1} - \delta\epsilon_{m+2} + \epsilon_{m+3} + \delta\epsilon_{m+4})\right|^2 \tag{17}$$

$$+ \frac{1}{T^2}\sum_{t,t' \bmod 4=1}\left|-2\delta - \epsilon_t - \epsilon_{t'} - \frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1} - \delta\epsilon_{m+2} + \epsilon_{m+3} + \delta\epsilon_{m+4})\right|^2 \tag{18}$$

$$+ \frac{1}{T^2}\sum_{t,t' \bmod 4=2}\left|2\delta - \delta\epsilon_t - \delta\epsilon_{t'} - \frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1} - \delta\epsilon_{m+2} + \epsilon_{m+3} + \delta\epsilon_{m+4})\right|^2 \tag{19}$$

$$+ \frac{1}{T^2}\sum_{t,t' \bmod 4=3}\left|-2\delta + \epsilon_t + \epsilon_{t'} - \frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1} - \delta\epsilon_{m+2} + \epsilon_{m+3} + \delta\epsilon_{m+4})\right|^2 \tag{20}$$

$$+ \frac{1}{T^2}\sum_{t \bmod 4=0,t' \bmod 4=2}\left|2\delta + \delta\epsilon_t - \delta\epsilon_{t'} - \frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1} - \delta\epsilon_{m+2} + \epsilon_{m+3} + \delta\epsilon_{m+4})\right|^2 \tag{21}$$

$$+\frac{1}{T^2}\sum_{t \bmod 4=2, t' \bmod 4=0}\left|2\delta-\delta\epsilon_t+\delta\epsilon_{t'}-\frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1}-\delta\epsilon_{m+2}+\epsilon_{m+3}+\delta\epsilon_{m+4})\right|^2 \tag{22}$$

$$+\frac{1}{T^2}\sum_{t \bmod 4=1, t' \bmod 4=3}\left|-2\delta-\epsilon_t+\epsilon_{t'}-\frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1}-\delta\epsilon_{m+2}+\epsilon_{m+3}+\delta\epsilon_{m+4})\right|^2 \tag{23}$$

$$+\frac{1}{T^2}\sum_{t \bmod 4=3, t' \bmod 4=1}\left|-2\delta+\epsilon_t-\epsilon_{t'}-\frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1}-\delta\epsilon_{m+2}+\epsilon_{m+3}+\delta\epsilon_{m+4})\right|^2 \tag{24}$$

$$\leq\frac{4}{T^2}\sum_{t,t' \bmod 4=0}\delta^2+\frac{1}{T^2}\sum_{t,t' \bmod 4=0}\left|\delta\epsilon_t+\delta\epsilon_{t'}-\frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1}-\delta\epsilon_{m+2}+\epsilon_{m+3}+\delta\epsilon_{m+4})\right|^2 \tag{25}$$

$$\cdots \tag{26}$$

$$+\frac{4}{T^2}\sum_{t \bmod 4=3, t' \bmod 4=1}\delta^2+\frac{1}{T^2}\sum_{t \bmod 4=3, t' \bmod 4=1}\left|\epsilon_t-\epsilon_{t'}-\frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1}-\delta\epsilon_{m+2}+\epsilon_{m+3}+\delta\epsilon_{m+4})\right|^2 \tag{27}$$

where in the last step we apply $|a+b|^2 \leq 2a^2+2b^2, \forall a, b \in \mathbb{R}$, and we break each term into two parts. The first part is independent of ϵ_t and is only related to δ^2, and the second term is the average of some sequence of ϵ_t. Then $\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\middle|\mathcal{E}\right]$ can be upper bounded by the following form

$$\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\middle|\mathcal{E}\right] \leq O(\delta^2)+\mathcal{T}_\epsilon, \tag{28}$$

where $\mathcal{T}_\epsilon$ is the term containing all the ϵ_t. Next we prove that $\mathcal{T}_\epsilon$ is a small quantity $o(\delta)$ with high probability. Note that here all the ϵ_t are i.i.d. random variables, by applying the Hoeffding's inequality, which is for i.i.d random variables $Z_1, \cdots, Z_N$ following Gaussian white noise distribution,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{m=1}^{N}Z_m\right|>t\right)\leq e^{-2Nt^2}, \tag{29}$$

applying this to $\mathcal{T}_\epsilon$, we obtain (we show the derivation on the first term, the others are similar), with probability $1-e^{-2To(\delta)^2}$,

$$\left|\delta\epsilon_t+\delta\epsilon_{t'}-\frac{2}{T}\sum_{m=0}^{\tilde{T}-1}(-\epsilon_{m+1}-\delta\epsilon_{m+2}+\epsilon_{m+3}+\delta\epsilon_{m+4})\right|^2 \tag{30}$$

$$\leq\frac{1}{4}\left|\frac{1}{\tilde{T}}\sum_{m=0}^{\tilde{T}-1}(\epsilon_{m+1}-\epsilon_{m+3})\right|^2+\frac{1}{4}\left|\frac{\delta}{\tilde{T}}\sum_{m=0}^{\tilde{T}-1}(\epsilon_{m+2}-\epsilon_{m+4}+\epsilon_t+\epsilon_{t'})\right|^2 \tag{31}$$

$$\leq o(\delta). \tag{32}$$

Apply this to every term, we obtain with high probability,

$$\mathcal{T}_\epsilon \leq o(\delta). \tag{33}$$

We can do the similar analysis on $\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\middle|\neg\mathcal{E}\right]$, and obtain $\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\middle|\neg\mathcal{E}\right] \leq o(\delta)$. Here we omit this part for brevity. □

B.2. Proof to Theorem 1

Proof. Since this theorem states the existence of a dataset in order to show a lower bound, the proof of is done by constructing such a dataset. This implies δ and N can be chosen freely. Without the loss of generality, in the proof we treat the mini-batch $M=1$.

When $p = 1$, we let $\mathcal{D}$ contains one single time series of length $p + 1$ and $\delta > 0$. It is straightforward to see the variance is zero (since the training here would be deterministic) and the theorem holds.

When $p \geq 2$, the $\mathcal{D}$ we construct contains $N = 2\lfloor p/2 \rfloor$ different time series, and each time series is of length $p + 1$. Note that here we have $N \geq 2$ because $p \geq 2$. Since the model is predicting the same time here (it is predicting time $p + 1$ given time 1 to p), we let $c = \epsilon_p$ denote a fixed value[6]. Without the loss of generality let $2c \leq \delta$ such that $\max_{i,t} |\boldsymbol{z}_{i,t}| = \delta$. Within the proof of this theorem, we let $\overline{p} = \lfloor p/2 \rfloor$, $\mathcal{D}$ is constructed as the following:

$$\text{Time Series i:} \quad \begin{bmatrix} \frac{\delta}{2}, \underbrace{0, \cdots, 0}_{p-1}, \frac{\delta}{2} + c \end{bmatrix}, i = 1$$

$$\begin{bmatrix} \underbrace{0, \cdots, 0}_{i-2}, \frac{\delta}{2}, -\frac{\delta}{2}, \underbrace{0, \cdots, 0}_{p-i}, c \end{bmatrix}, 2 \leq i \leq \overline{p}$$

$$\begin{bmatrix} \underbrace{0, \cdots, 0}_{\overline{p}}, \underbrace{\frac{\delta}{2} \cdots, \frac{\delta}{2}}_{p-\overline{p}}, \frac{\delta}{2} + c \end{bmatrix}, \overline{p} + 1 \leq i \leq 2\overline{p}.$$

Fit in the MSE loss we obtain:

$$f_1(\boldsymbol{\theta}) = \left(\frac{\delta}{2} - \frac{\delta}{2} \boldsymbol{\theta}_1 \right)^2$$

$$f_i(\boldsymbol{\theta}) = \left(\frac{\delta}{2} \boldsymbol{\theta}_{i-1} - \frac{\delta}{2} \boldsymbol{\theta}_i \right)^2, \forall 2 \leq i \leq \overline{p}$$

$$f_i(\boldsymbol{\theta}) = \left(\frac{\delta}{2} - \frac{\delta}{2} \boldsymbol{\theta}_{\overline{p}+1} - \cdots - \frac{\delta}{2} \boldsymbol{\theta}_p \right)^2, \forall \overline{p} + 1 \leq i \leq 2\overline{p},$$

where f_i denotes the loss incurred on the i-th time series. The total loss function can then be expressed as

$$f(\boldsymbol{\theta}) = \frac{1}{2\overline{p}} \sum_{i=1}^{2\overline{p}} f_i(\boldsymbol{\theta}) = \underbrace{\frac{1}{2\overline{p}} \sum_{i=1}^{\overline{p}} f_i(\boldsymbol{\theta})}_{g_1(\boldsymbol{\theta})} + \underbrace{\frac{1}{2\overline{p}} \sum_{i=\overline{p}+1}^{2\overline{p}} f_i(\boldsymbol{\theta})}_{g_2(\boldsymbol{\theta})}. \tag{34}$$

Note that when taking derivative of function f with respect to the $\boldsymbol{\theta}$, the first $\overline{p}$ coordinates will only be affected by $g_1(\boldsymbol{\theta})$, i.e.,

$$\frac{\partial f}{\partial \boldsymbol{\theta}_i} = \frac{\partial g_1}{\partial \boldsymbol{\theta}_i}, \forall i \leq \overline{p}, \tag{35}$$

then we obtain

$$\|\nabla f(\boldsymbol{\theta})\|^2 = \sum_{j=1}^{p} \left| \frac{\partial f}{\partial \boldsymbol{\theta}_i} \right|^2 \geq \sum_{j=1}^{\overline{p}} \left| \frac{\partial f}{\partial \boldsymbol{\theta}_i} \right|^2 = \sum_{j=1}^{\overline{p}} \left| \frac{\partial g_1}{\partial \boldsymbol{\theta}_i} \right|^2 = \|\nabla g_1(\boldsymbol{\theta})\|^2. \tag{36}$$

Lemma 1 in (Carmon et al., 2019) shows that for every $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_{\overline{p}} = 0$,

$$\|\nabla g_1(\boldsymbol{\theta})\| \geq \frac{\delta^2}{4\overline{p}^{\frac{5}{2}}}. \tag{37}$$

As a result, we obtain for every $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_{\overline{p}} = 0$,

$$\|\nabla f(\boldsymbol{\theta})\| \geq \frac{\delta^2}{4\overline{p}^{\frac{5}{2}}}. \tag{38}$$

Without the loss of generality we set[7] $\boldsymbol{\theta}^{(0)} = \mathbf{0}$. Now if we look at the expression of function g_1, if the model starts from $\mathbf{0}$,

[6] The ϵ_p can be obtained by generating the dataset using the same random seed as used in the model.

[7] If the initialization $\boldsymbol{v} \neq \mathbf{0}$, we just need to replace all the $\boldsymbol{\theta}$ with $\boldsymbol{\theta} - \boldsymbol{v}$ in the original functions and the proof will be the same.

it follows a "zero-respecting" property: $\boldsymbol{\theta}_j$ will remain zero if fewer then j number of gradients on g_1 is computed (Carmon et al., 2019). Define a filtration $\mathcal{F}^{(t)}$ at iteration t as the sigma field of all the previous events happened before iteration t. Let τ_i denote the recent time where sample i is sampled for computing the stochastic gradient. And let N_t be a random variable, denoting the largest number where τ_1 to τ_{N_t} is strictly increasing. Since each sample is uniformly sampled, we obtain

$$\mathbb{P}[N^{(t+1)} - N^{(t)} = 1 | \mathcal{F}^{(t)}] \leq \frac{1}{2\overline{p}} \leq \frac{1}{\overline{p}}. \tag{39}$$

Let $q^{(t)} = N^{(t+1)} - N^{(t)}$, with Chernoff bound, we obtain

$$\mathbb{P}[N^{(t)} \geq \overline{p}] = \mathbb{P}[e^{\sum_{j=0}^{t-1} q^{(j)}} \geq e^{\overline{p}}] \leq e^{-\overline{p}} \mathbb{E}[e^{\sum_{j=0}^{t-1} q^{(j)}}]. \tag{40}$$

For the expectation term we know that

$$\mathbb{E}[e^{\sum_{j=0}^{t-1} q^{(j)}}] = \mathbb{E}\left[\prod_{j=0}^{t-1} \mathbb{E}\left[e^{q^{(j)}} | \mathcal{F}^{(j)}\right]\right] \leq \left(1 - \frac{1}{\overline{p}} + \frac{e}{\overline{p}}\right)^t \leq e^{t(e-1)/\overline{p}}. \tag{41}$$

Thus we know

$$\mathbb{P}[N^{(t)} \geq \overline{p}] \leq e^{\frac{(e-1)t}{\overline{p}} - \overline{p}} \leq \omega, \tag{42}$$

for every $t \leq \frac{\overline{p}^2 + \overline{p}\log(\omega)}{(e-1)}$. Take $\omega = \frac{1}{2}$, for any $0 < \epsilon < \frac{\delta^2}{8p^{\frac{5}{2}}} \leq \frac{\delta^2}{8\overline{p}^{\frac{5}{2}}}$, we obtain

$$\begin{aligned}
\mathbb{E}\|\nabla f(\boldsymbol{\theta})\| =& \mathbb{P}(N^{(t)} \geq \overline{p}) \left[\|\nabla f(\boldsymbol{\theta})\| \middle| N^{(t)} \geq \overline{p}\right] + \mathbb{P}(N^{(t)} < \overline{p}) \left[\|\nabla f(\boldsymbol{\theta})\| \middle| N^{(t)} < \overline{p}\right] && (43)\\
\geq& \frac{1}{2}\|\nabla f(\boldsymbol{\theta})\| && (44)\\
>& \frac{1}{2} \cdot 2\epsilon && (45)\\
=& \epsilon, && (46)
\end{aligned}$$

where we use Equation (38). The gradient is calculated as follows:

$$\begin{aligned}
\nabla f_1(\mathbf{0}) &= \left[-\frac{\delta^2}{2}, \underbrace{0, \cdots, 0}_{p-1}\right]\\
\nabla f_i(\mathbf{0}) &= \left[\underbrace{0, \cdots, 0}_{p}\right], 2 \leq i \leq \overline{p}\\
\nabla f_i(\mathbf{0}) &= \left[\underbrace{0, \cdots, 0}_{\overline{p}}, \underbrace{-\frac{\delta^2}{2} \cdots, -\frac{\delta^2}{2}}_{p-\overline{p}}\right], \overline{p}+1 \leq i \leq 2\overline{p}\\
\nabla f(\mathbf{0}) &= \left[-\frac{\delta^2}{4\overline{p}}, \underbrace{0, \cdots, 0}_{\overline{p}-1}, \underbrace{-\frac{\delta^2}{4\overline{p}} \cdots, -\frac{\delta^2}{4\overline{p}}}_{p-\overline{p}}\right].
\end{aligned}$$

The sampling variance in the first iteration:

$$\begin{aligned}
\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\right] =& \mathbb{E}_{i \sim [2\overline{p}]} \|\nabla f_i(\mathbf{0}) - \nabla f(\mathbf{0})\|^2 && (47)\\
=& \frac{1}{2\overline{p}} \left[\left(\frac{2\overline{p}-1}{4\overline{p}}\right)^2 \delta^4 + \frac{p-\overline{p}}{16\overline{p}^2}\delta^4\right] + \frac{\overline{p}-1}{2\overline{p}}\left[\frac{p-\overline{p}+1}{16\overline{p}^2}\delta^4\right] + \frac{\overline{p}}{2\overline{p}}\left[\frac{1}{16\overline{p}^2}\delta^4 + (p-\overline{p})\left(\frac{2\overline{p}-1}{4\overline{p}}\right)^2 \delta^4\right] && (48)
\end{aligned}$$

$$\leq \frac{3}{\overline{p}}\delta^4 + \overline{p}\delta^4 \tag{49}$$

$$\leq \overline{p}^2, \tag{50}$$

where the last step holds when we let $\delta^4 \leq \min\{\overline{p}^3/6, \overline{p}\}$. Since for every $t \leq \frac{\overline{p}^2+\overline{p}\log(\omega)}{(e-1)}$, $\mathbb{E}\|\nabla f(\boldsymbol{\theta})\| \geq \epsilon$, the lower bound on the iterations (number of gradients to be computed) T_b is

$$T_b = \Omega\left(\overline{p}^2\right), \tag{51}$$

which implies

$$T_b = \Omega\left(\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\right]\right). \tag{52}$$

Furthermore,

$$\text{Var}\left[\nabla f_{\xi^{(0)}}(\boldsymbol{\theta}^{(0)})\right] = \frac{1}{2\overline{p}}\left[\left(\frac{2\overline{p}-1}{4\overline{p}}\right)^2\delta^4 + \frac{p-\overline{p}}{16\overline{p}^2}\delta^4\right] + \frac{\overline{p}-1}{2\overline{p}}\left[\frac{p-\overline{p}+1}{16\overline{p}^2}\delta^4\right] + \frac{\overline{p}}{2\overline{p}}\left[\frac{1}{16\overline{p}^2}\delta^4 + (p-\overline{p})\left(\frac{2\overline{p}-1}{4\overline{p}}\right)^2\delta^4\right] \tag{53}$$

$$\geq \frac{p-\overline{p}}{2}\left(\frac{2\overline{p}-1}{4\overline{p}}\right)^2\delta^4 \tag{54}$$

$$= \Omega(\delta^4 p), \tag{55}$$

and thus we complete the proof. □

B.3. Proof to Theorem 2

B.3.1. Main Proof

Proof. Take the expectation with respect to the sampling randomness in the inner loop for $\boldsymbol{v}^{(t,k)}$, we obtain

$$\mathbb{E}_{\xi^{(t,k)}}\left[\boldsymbol{v}^{(t,k)}\right] = \mathbb{E}_{\xi^{(t,k)}}\left[\nabla f(\boldsymbol{\theta}^{(t,k)};\xi^{(t,k)}) - \nabla f(\boldsymbol{\theta}^{(t,0)};\xi^{(t,k)}) + \boldsymbol{g}^{(t,0)}\right] = \nabla f(\boldsymbol{\theta}^{(t,k)}) + \underbrace{\boldsymbol{g}^{(t,0)} - \nabla f(\boldsymbol{\theta}^{(t,0)})}_{=\boldsymbol{\zeta}_t} \tag{56}$$

Due to $\boldsymbol{\zeta}_t$, the main step for SCott ($\boldsymbol{v}^{(t,k)}$) is a biased estimation of the true gradient $\nabla f\left(\boldsymbol{\theta}^{(t,k)}\right)$. The challenge of the proof is to handle such biasedness. The rest of our analysis largely follows the proof routine in SCSG-type methods (Babanezhad Harikandeh et al., 2015; Lei et al., 2017; Li & Li, 2018). We do not take credit for those analysis.

Li & Li (2018) proposes a nice framework of analyzing stochastic control variate type algorithm, where in their framework, the control variate $\boldsymbol{g}^{(t,0)}$ is computed via a randomly sampled mini-batch. This, as we discussed in the paper, can be seen as a special case of random grouping. The main difference of SCott is in bounding $\boldsymbol{\zeta}_t$ since here the noise is not from the uniform sampling. For brevity, we summarize several lemmas from previous work. and focus on analyzing $\boldsymbol{\zeta}_t$ in the main proof.

We summarize the main results in Lemma 4. We encourage readers to refer to (Li & Li, 2018) for complete derivation for this part of results. From Lemma 4 we obtain when $\alpha_t L = cB^{-\frac{2}{3}}$ (where c is a numerical constant),

$$\alpha_t B\left(2 - \frac{2}{B} - 2\alpha_t L - \frac{1}{1-\alpha_t^2L^2B - \alpha_t^3L^3B^2}\right)\mathbb{E}\|\nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \tag{57}$$

$$\leq 2\mathbb{E}(f(\boldsymbol{\theta}^{(t-1,0)}) - f(\boldsymbol{\theta}^{(t,0)})) + 2\alpha_t B\left(1 + \alpha_t L + \frac{1}{B}\right)\mathbb{E}\|\boldsymbol{\zeta}_t\|^2, \tag{58}$$

given $\alpha_t L = cB^{-\frac{2}{3}}$, we obtain

$$1 - \alpha_t^2L^2B - \alpha_t^3L^3B^2 \geq 1 - B^{-\frac{1}{3}}c^2 - c^3, \tag{59}$$

put it back together with $\alpha_t L = cB^{-\frac{2}{3}}$, we get,

$$cB^{\frac{1}{3}}\left(2-\frac{2}{B}-2cB^{-\frac{2}{3}}-\frac{1}{1-B^{-\frac{1}{3}}c^2-c^3}\right)\mathbb{E}\|\nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \tag{60}$$

$$\leq 2L\mathbb{E}(f(\boldsymbol{\theta}^{(t-1,0)})-f(\boldsymbol{\theta}^{(t,0)}))+2cB^{\frac{1}{3}}\left(1+cB^{-\frac{2}{3}}+B^{-1}\right)\mathbb{E}\|\boldsymbol{\zeta}_t\|^2, \tag{61}$$

select $c \leq \frac{1}{4}$ we can get,

$$2-\frac{2}{B}-2cB^{-\frac{2}{3}}-\frac{1}{1-B^{-\frac{1}{3}}c^2-c^3} \geq \frac{1}{4} \tag{62}$$

$$1+cB^{-\frac{2}{3}}+B^{-1} \leq 1.35. \tag{63}$$

Fit in Lemma 5, we obtain

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \leq \frac{8L\mathbb{E}(f(\boldsymbol{\theta}^{(t-1,0)})-f(\boldsymbol{\theta}^{(t,0)}))}{cB^{\frac{1}{3}}}+11\sum_{i=1}^{B}w_i^2\sigma_i^2. \tag{64}$$

Telescoping from $t=0$ to $T-1$, we obtain

$$\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{\theta}}^{(T)}\right)\right\|^2 \leq \frac{8(f(\mathbf{0})-\inf_{\boldsymbol{\theta}} f(\boldsymbol{\theta}))L}{cB^{\frac{1}{3}}T}+11\sum_{i=1}^{B}w_i^2\sigma_i^2 \tag{65}$$

$$= O\left(\frac{(f(\mathbf{0})-\inf_{\boldsymbol{\theta}} f(\boldsymbol{\theta}))L}{B^{\frac{1}{3}}T}+\sum_{i=1}^{B}w_i^2\sigma_i^2\right), \tag{66}$$

given our selection on the value of B,

$$\sum_{i=1}^{B}w_i^2\sigma_i^2 \leq O\left(\epsilon^2\right), \tag{67}$$

and then, with

$$T = O\left(\frac{(f(\mathbf{0})-\inf_{\boldsymbol{\theta}} f(\boldsymbol{\theta}))L}{B^{\frac{1}{3}}\epsilon^2}\right), \tag{68}$$

we obtain

$$\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{\theta}}^{(T)}\right)\right\| \leq \sqrt{\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{\theta}}^{(T)}\right)\right\|^2} \tag{69}$$

$$= \sqrt{O\left(\frac{(f(\mathbf{0})-\inf_{\boldsymbol{\theta}} f(\boldsymbol{\theta}))L}{B^{\frac{1}{3}}T}+\sum_{i=1}^{B}w_i^2\sigma_i^2\right)} \tag{70}$$

$$\leq \epsilon. \tag{71}$$

Applying Lemma 3, the total number of stochastic gradient being computed can then be calculated as

$$\sum_{t=0}^{T-1}(B+\mathbb{E}[K_t]) = 2BT = O\left(\frac{\Delta L B^{\frac{2}{3}}}{\epsilon^2}\right), \tag{72}$$

when $B=|\mathcal{D}|$, then $\sigma^2_{|\mathcal{D}|}=0$, the total number of gradients to be computed is

$$O\left(\frac{\Delta L|\mathcal{D}|^{\frac{2}{3}}}{\epsilon^2}\right), \tag{73}$$

when $B \neq |\mathcal{D}|$, then put in

$$B = O\left(\frac{B\sum_{i=1}^{B} w_i^2\sigma_i^2}{\epsilon^2}\right), \tag{74}$$

the total number of gradients to be computed is

$$O\left(\frac{\Delta L\left(B\sum_{i=1}^{B} w_i^2\sigma_i^2\right)^{\frac{2}{3}}}{\epsilon^{\frac{10}{3}}}\right). \tag{75}$$

And thus we complete the proof. □

B.3.2. Technical Lemma

Lemma 3. *((Horváth et al., 2020), Lemma 1) Let $N \sim Geom(\gamma)$, for $\gamma > 0$. Then for any sequence $D_0, D_1, \cdots$ with $\mathbb{E}|D_N| \leq \infty$,*

$$\mathbb{E}(D_N - D_{N+1}) = \left(\frac{1}{\gamma} - 1\right)(D_0 - \mathbb{E}D_N). \tag{76}$$

Lemma 4. *((Li & Li, 2018), Proof to Theorem 3.1) when $\alpha_t L = cB^{-\frac{2}{3}}$,*

$$\alpha_t B\left(2 - \frac{2}{B} - 2\alpha_t L - \frac{1}{1-\alpha_t^2L^2B - \alpha_t^3L^3B^2}\right)\mathbb{E}\|\nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \tag{77}$$

$$\leq 2\mathbb{E}(f(\boldsymbol{\theta}^{(t-1,0)}) - f(\boldsymbol{\theta}^{(t,0)})) + 2\alpha_t B\left(1 + \alpha_t L + \frac{1}{B}\right)\mathbb{E}\|\boldsymbol{\zeta}_t\|^2. \tag{78}$$

Proof. The proof can be established straightforwardly by considering the contant batch size case in (Li & Li, 2018). □

Lemma 5.

$$\mathbb{E}\|\boldsymbol{\zeta}_t\|^2 \leq \sum_{i=1}^{B} w_i^2\sigma_i^2. \tag{79}$$

Proof.

$$\mathbb{E}\|\boldsymbol{\zeta}_t\|^2 \tag{80}$$

$$= \mathbb{E}\|\boldsymbol{g}^{(t,0)} - \nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \tag{81}$$

$$= \mathbb{E}\left\|\sum_{i=1}^{B}\frac{|\mathcal{D}_i|\nabla f(\boldsymbol{\theta}^{(t,0)};\xi_i^{(t)})}{|\mathcal{D}|} - \nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 \tag{82}$$

$$= \mathbb{E}\left\|\sum_{i=1}^{B}\frac{|\mathcal{D}_i|\nabla f(\boldsymbol{\theta}^{(t,0)};\xi_i^{(t)})}{|\mathcal{D}|} - \sum_{i=1}^{B}\frac{|\mathcal{D}_i|}{|\mathcal{D}|}\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 \tag{83}$$

$$= \sum_{i=1}^{B} w_i^2\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)};\xi_i^{(t)}) - \nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 + \sum_{i\neq i'}\mathbb{E}\langle\nabla f(\boldsymbol{\theta}^{(t,0)};\xi_i^{(t)}) - \nabla f(\boldsymbol{\theta}^{(t,0)}), \nabla f(\boldsymbol{\theta}^{(t,0)};\xi_{i'}^{(t)}) - \nabla f(\boldsymbol{\theta}^{(t,0)})\rangle \tag{84}$$

$$\leq \sum_{i=1}^{B} w_i^2\sigma_i^2 \tag{85}$$

where in the final step, we use the property that $\xi_i^{(t)}$ is independent of $\xi_{i'}^{(t)}$. □