Supplementary Material for GR-Gauge: Cost-efficient Training Configuration By Gauging the Gradient Redundancy

Guanjie Wang
Zhiyuan College, Shanghai Jiao Tong University

Chen Chen
Shanghai Jiao Tong University

A. Assumptions for Theoretical Analysis

To model the metrics and establish the theoretical analysis, we make the following assumptions. Table 1 summarizes the key notations used in the analysis.

Table 1. Key notations used in the analysis.

$g_x(\theta)$	The gradient computed on a data point x at the parameter θ
$G_{est}(\theta)$	The sampled batch gradient of the overall loss, i.e., $G_{est}(\theta) = \frac{1}{B}\sum_{i=1}^{B} g_{x_i}(\theta), x_i \sim \rho$
$G(\theta)$	The gradient of the overall loss, i.e., $G(\theta) = \mathbb{E}_{x\sim\rho}[g_x(\theta)]$
H	The constant Hessian matrix of the overall loss function
μ	The constant scale that quantifies the gradient noise variance, i.e., $\mu = Tr(Cov_{x\sim\rho}[g_x(\theta)])/\|\|G(\theta)\|\|_2^2$
β	The constant smooth factor of the exponential moving average used to calculate GR_T, which is fixed to 0.9 in the paper
η	The learning rate
B	The global batch size
$\bar{\lambda}$	The average eigenvalue of H

Assumption 1. ***(Gradient Distribution)*** *Assume that for any parameter θ, the expectation and variance of stochastic gradient $g_x(\theta)$ computed on a data point x satisfy that:*

$$\mathbb{E}_{x\sim\rho}[g_x(\theta)] = G(\theta), \quad Tr(Cov_{x\sim\rho}(g_x(\theta))) = \mu(\theta)||G(\theta)||_2^2,$$

where ρ is the data distribution, $\mu(\theta)$ is the scalar that reflects the scale of gradient noise variance introduced by the dataset.

Assumption 2. ***(Local Smoothness and Stability)*** *Within a small neighborhood of θ (short-term dynamics), we assume that the Hessian matrix $H(\theta)$ of the overall loss function and gradient noise scale $\mu(\theta)$ can be approximated as constants:*

$$H(\theta) \approx H, \quad \mu(\theta) \approx \mu.$$

Assumption 3. ***(Independent and Identically Distributed Gradients)*** *The stochastic gradients $\{g_{x_i}(\theta)\}_{i=1}^N$ computed on different data samples $\{x_i\}_{i=1}^N$ are independent and identically distributed (i.i.d.), satisfying:*

$$g_{x_i}(\theta) \overset{i.i.d.}{\sim} \rho_\theta \quad \textit{for all } i = 1, ..., N,$$

where ρ_θ is a distribution characterized by Assumption 1, N is the dataset size.

B. Proof of Theorem 1

Before proving Theorem 1, we first prove Lemma 1.

Lemma 1. *Under the Assumption 1, Assumption 2 and Assumption 3, denote the learning rate by η and the average eigenvalue of the Hessian matrix H of the loss function by $\bar{\lambda}$. Denote the global batch size by B and the gradient noise scale by μ. β is the constant smooth factor defined in Definition 1. The theoretical expected stable value of GR_T is*

$$\mathbb{E}[GR_T] \approx \frac{\left((1+\frac{\mu}{B})^2(1-\bar{\lambda}\eta-\beta)^2\right)\left(\beta^{\frac{1}{1-\beta}} - \left(1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}\right)^{\frac{1}{1-\beta}}\right)}{(1-\beta)\left(2\bar{\lambda}\eta-(1-\beta)\left(1+\frac{\mu}{B}\right)\right)\left(\beta^{\frac{1}{1-\beta}} - \left(1-\bar{\lambda}\eta\right)^{\frac{1}{1-\beta}}\right)^2}. \tag{1}$$

Proof. For a given dataset with data points in the distribution $x \sim D$, each data point x at parameter θ is associated with the gradient $g_x(\theta)$. Denote the global batch size as B. Then the sampled batch gradient $G_{est}(\theta)$ at parameter θ is

$$G_{est}(\theta) = \frac{1}{B}\sum_{i=1}^{B} g_{x_i}(\theta); \quad x_i \sim \rho. \tag{2}$$

During the training process, denote the learning rate as η, and the parameter at the t-th iteration be θ_t. It can be considered that

$$\theta_{t+1} = \theta_t - \eta G_{est}(\theta_t). \tag{3}$$

Since the update can be regarded as a small quantity, we perform a Taylor expansion as

$$G_{est}(\theta_{t+1}) = \frac{1}{B}\sum_{i=1}^{B} g_{x_i}(\theta_t - \eta G_{est}(\theta_t)) \approx \frac{1}{B}\sum_{i=1}^{B} g_{x_i}(\theta_t) - \frac{\eta}{B}\sum_{i=1}^{B} H_{x_i}(\theta_t)^T G_{est}(\theta_t), \tag{4}$$

where $H_{x_i}(\theta_t)$ denotes the Hessian matrix of the loss function on data point x_i at parameter θ_t. Since the update is regarded as a small quantity and Assumption 2, the Hessian matrix of the different loss function and parameters can be regarded as constant, which is denoted as H. Then we take the expectation of Eq. 2 and Eq. 4 and obtain that

$$\mathbb{E}_{x\sim\rho}[G_{est}(\theta_t)] = G(\theta_t), \tag{5}$$

$$\mathbb{E}_{x\sim\rho}[G_{est}(\theta_{t+1})] = (I - \eta H^T)\mathbb{E}_{x\sim\rho}[G_{est}(\theta_t)]. \tag{6}$$

Adopting the idea of chunking, we set $t = t_0$ as the new time zero point. Let the parameter at new iteration 0 be θ_0. Therefore, at the new iteration t, we have

$$\mathbb{E}_{x\sim\rho}[G_{est}(\theta_t)] = (I - \eta H)^t G(\theta_0). \tag{7}$$

Then we calculate the theoretical expected value of $||\hat{m}_t||_2^2$ which is defined in definition 1. By Eq. 7, it's obvious that

$$\begin{aligned}\mathbb{E}_{x\sim\rho}[\hat{m}_t] \approx& \beta^i \cdot \mathbb{E}_{x\sim\rho}[\hat{m}_0] + (1-\beta)\sum_{k=0}^{t-1}\beta^{t-1-k}\cdot\mathbb{E}_{x\sim\rho}[G_{est}(\theta_k)] \\ =& \beta^i \cdot \mathbb{E}_{x\sim\rho}[\hat{m}_0] + (1-\beta)\beta^{t-1}\left(\sum_{k=0}^{t-1}\left(\frac{I-\eta H}{\beta}\right)^k\right)G(\theta_0).\end{aligned} \tag{8}$$

Assume the eigenvalue decomposition of H is given by $H = P\Lambda P^T$, where P is the orthogonal matrix of eigenvectors and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues. Denote the average eigenvalue of H as $\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i$. Then by transformation and approximate, we know that (we leverage the formula $P = P^T = P^{-1}$ when P is orthogonal matrix, and ignore $\hat{m}_0$ since β^t can be regarded as a small value as t increasing)

$$\begin{aligned}||\mathbb{E}_{x\sim\rho}[\hat{m}_t]||_2^2 &\approx (1-\beta)^2\beta^{2t-2}G^T(\theta_0)P\left(\sum_{k=0}^{t-1}\left(\frac{I-\eta\Lambda}{\beta}\right)^k\right)^2 P^T G(\theta_0) \\ &\approx (1-\beta)^2\beta^{2t-2}\left(\sum_{k=0}^{t-1}\left(\frac{1-\bar{\lambda}\eta}{\beta}\right)^k\right)^2 G^T(\theta_0)PP^T G(\theta_0) \\ &= \frac{(1-\beta)^2\beta^{2t}}{(1-\bar{\lambda}\eta-\beta)^2}\left(\left(\frac{1-\bar{\lambda}\eta}{\beta}\right)^t - 1\right)^2 ||G(\theta_0)||_2^2.\end{aligned} \tag{9}$$

Next, we analyze the variance. Since the batches selected between iteration t and $t+1$ are independent (which are denoted as

x_i and x_j), we can get that (ignoring higher-order terms)

$$
\begin{aligned}
&G_{est}(\theta_{t+1})G_{est}^T(\theta_{t+1}) \\
\approx&\left(\frac{1}{B}\sum_{i=1}^{B}g_{x_i}(\theta_t)-\frac{\eta H^T}{B}\sum_{j=1}^{B}g_{x_j}(\theta_t)\right)\left(\frac{1}{B}\sum_{i=1}^{B}g_{x_i}(\theta_t)-\frac{\eta H^T}{B}\sum_{j=1}^{B}g_{x_j}(\theta_t)\right)^T \\
\approx&\frac{1}{B^2}\sum_{i=1}^{B}g_{x_i}(\theta_t)\sum_{i=1}^{B}g_{x_i}^T(\theta_t)-\frac{\eta H^T}{B^2}\sum_{j=1}^{B}g_{x_j}(\theta_t)\sum_{i=1}^{B}g_{x_i}^T(\theta_t)-\frac{\eta}{B^2}\sum_{i=1}^{B}g_{x_i}(\theta_t)\sum_{j=1}^{B}g_{x_j}^T(\theta_t)H.
\end{aligned}
\tag{10}
$$

Then by Assumption 3, we take the expectation

$$
\begin{aligned}
&\mathbb{E}_{x\sim\rho}\left[G_{est}(\theta_{t+1})G_{est}^T(\theta_{t+1})\right] \\
\approx&\left(\frac{1}{B}\mathbb{E}_{x\sim\rho}[g_x(\theta)g_x^T(\theta)]+\frac{B-1}{B}G(\theta)G^T(\theta)\right)-\eta H^TG(\theta_t)G^T(\theta_t)-\eta G(\theta_t)G^T(\theta_t)H.
\end{aligned}
\tag{11}
$$

Since the covariance matrix is defined as $Cov_{x\sim\rho}(g_x(\theta))=\mathbb{E}_{x\sim\rho}[g_x(\theta)g_x^T(\theta)]-G(\theta)G^T(\theta)$, we can obtain that

$$
Tr\left(\mathbb{E}_{x\sim\rho}\left[G_{est}(\theta_{t+1})G_{est}^T(\theta_{t+1})\right]\right)\approx\frac{1}{B}Tr\left(Cov_{x\sim\rho}(g_x(\theta))\right)+Tr\left(G(\theta_t)G^T(\theta_t)\right)-2\eta Tr\left(G(\theta_t)G^T(\theta_t)H\right). \tag{12}
$$

Similarly, we decompose $H=P\Lambda P^T$, and obtain the results by transformation and approximate (we leverage the formulas $Tr(AB)=Tr(BA)$ and $Tr(PAP^T)=Tr(A)$, where P is the orthogonal matrix)

$$
Tr\left(G(\theta_t)G^T(\theta_t)H\right)=Tr\left(G(\theta_t)G^T(\theta_t)P\Lambda P^T\right)=Tr\left(\Lambda P^T\left(G(\theta_t)G^T(\theta_t)\right)P\right)\approx\bar{\lambda}Tr\left(G(\theta_t)G^T(\theta_t)\right). \tag{13}
$$

From Assumption 1 and Assumption 2 we know that $Tr(Cov_{x\sim\rho}(g_x(\theta)))=\mu||G(\theta_t)||^2$, where μ is the constant gradient noise scale. Then based on Eq. 12, Eq. 13 and assumptions we have (we leverage the formulas $Tr(xx^T)=||x||_2^2, x\in R^d$ and $\mathbb{E}[Tr(A)]=Tr\left(\mathbb{E}[A]\right)$)

$$
\mathbb{E}_{x\sim\rho}\left[||G_{est}(\theta_{t+1})||_2^2\right]\approx\left(1-2\bar{\lambda}\eta+\frac{\mu}{B}\right)||G(\theta_t)||_2^2. \tag{14}
$$

Since Eq. 14 holds for any $\bar{\lambda}\eta$, we let $\bar{\lambda}\eta\to 0$ and have $\theta_{t+1}\to\theta_t$. Then we obtain that

$$
\mathbb{E}_{x\sim\rho}\left[||G_{est}(\theta_t)||_2^2\right]\approx\left(1+\frac{\mu}{B}\right)||G(\theta_t)||_2^2. \tag{15}
$$

Similarly, we have

$$
\mathbb{E}_{x\sim\rho}[||G_{est}(\theta_t)||_2^2]=\left(1+\frac{\mu}{B}\right)\left(1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}\right)^t||G(\theta_0)||_2^2. \tag{16}
$$

Also we calculate the theoretical expected value of $||\hat{v}_t||_1$ which is defined in definition 1. By Eq. 16, we can similarly get the result that (also ignore $\hat{v}_0$ due to the same reason)

$$
\begin{aligned}
\mathbb{E}_{x\sim\rho}[||\hat{v}_t||_1]&\approx\beta^i\cdot\mathbb{E}_{x\sim\rho}[||\hat{v}_0||_1]+(1-\beta)\sum_{k=0}^{t-1}\beta^{t-1-k}\mathbb{E}_{x\sim\rho}[||G_{est}(\theta_k)||_2^2] \\
&\approx\left(1+\frac{\mu}{B}\right)(1-\beta)\beta^{t-1}\left(\sum_{k=0}^{t-1}\left(\frac{1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}}{\beta}\right)^k\right)||G(\theta_0)||_2^2 \\
&=\frac{\left(1+\frac{\mu}{B}\right)^2(1-\beta)\beta^t}{(1-\beta)\left(1+\frac{\mu}{B}\right)-2\bar{\lambda}\eta}\left(\left(\frac{1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}}{\beta}\right)^t-1\right)||G(\theta_0)||_2^2.
\end{aligned}
\tag{17}
$$

Based on Eq. 9 and Eq. 17, we can get the theoretical expected value of $GR_T(t)$:

$$\mathbb{E}_{x\sim\rho}[GR_T(t)] \approx \frac{\mathbb{E}_{x\sim\rho}[||\hat{v}_t||_1]}{\mathbb{E}_{x\sim\rho}[||\hat{m}_t||_2^2]} \approx \frac{(1+\frac{\mu}{B})^2(1-\bar{\lambda}\eta-\beta)^2}{(1-\beta)\left(2\bar{\lambda}\eta-(1-\beta)\left(1+\frac{\mu}{B}\right)\right)} \cdot \frac{\beta^t-\left(1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}\right)^t}{\left(\beta^t-\left(1-\bar{\lambda}\eta\right)^t\right)^2}. \tag{18}$$

For t, because EMA processing is used, $t=\frac{1}{1-\beta}$ is used as the estimated time for the expected stable value of GR_T. Thus we obtain that

$$\mathbb{E}_{x\sim\rho}[GR_T] \approx \frac{(1+\frac{\mu}{B})^2(1-\bar{\lambda}\eta-\beta)^2}{(1-\beta)\left(2\bar{\lambda}\eta-(1-\beta)\left(1+\frac{\mu}{B}\right)\right)} \cdot \frac{\beta^{\frac{1}{1-\beta}}-\left(1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}\right)^{\frac{1}{1-\beta}}}{\left(\beta^{\frac{1}{1-\beta}}-\left(1-\bar{\lambda}\eta\right)^{\frac{1}{1-\beta}}\right)^2}. \tag{19}$$

Here, we complete the proof of Lemma 1. □

Then we proceed to prove Theorem 1.

Theorem 1. *There exists a positive partial derivative relationship between the learning rate η and the temporal gradient redundancy GR_T:*

$$\frac{\partial}{\partial\eta}\mathbb{E}[GR_T] > 0. \tag{20}$$

Proof. Based on Lemma 1, we perform a Taylor expansion of GR_T with respect to $\bar{\lambda}\eta$ and approximating to the first order

$$\mathbb{E}_{x\sim\rho}[GR_T] \approx \frac{1+\frac{\mu}{B}}{1-\beta^t} + \frac{2\frac{\mu}{B}}{1-\beta^t}\left(\frac{\frac{1}{1-\beta}}{1-\beta^{\frac{1}{1-\beta}}}-\frac{1}{1-\beta}\right)\bar{\lambda}\eta. \tag{21}$$

Therefore we have

$$\frac{\partial}{\partial\bar{\lambda}\eta}\mathbb{E}_{x\sim\rho}[GR_T] \approx \frac{2\frac{\mu}{B}\bar{\lambda}}{1-\beta^t}\left(\frac{\frac{1}{1-\beta}}{1-\beta^{\frac{1}{1-\beta}}}-\frac{1}{1-\beta}\right). \tag{22}$$

Since $0<\beta<1$, for $t\in\mathbb{N}_+$, the following inequality holds

$$\frac{i}{1-\beta^t}-\frac{1}{1-\beta} = \frac{t-(1+\beta+\cdots+\beta^{t-1})}{(1-\beta)(1+\beta+\cdots+\beta^{t-1})} > 0. \tag{23}$$

Through monotonicity, we know the the inequality holds when $i=\frac{1}{1-\beta}$. Now we obtain that

$$\frac{\partial}{\partial\eta}\mathbb{E}_{x\sim\rho}[GR_T] > 0. \tag{24}$$

Here, we complete the proof of Theorem 1. □

C. Proof of Theorem 2

Before proving Theorem 2, we first prove Lemma 2.

Lemma 2. *Under the Assumption 1, Assumption 2 and Assumption 3, denote the global batch size by B and the gradient noise scale by μ. The theoretical expected stable value of GR_S is*

$$\mathbb{E}\left[GR_S\right] \approx \frac{1+\mu}{B+\mu}. \tag{25}$$

Proof. By Assumption 1 and Assumption 2, we have (we leverage the formula $\mathbb{E}[||x^2||_1] = Tr(Cov_{x\sim\rho}(x))$ for $x \in R^d$):

$$\mathbb{E}_{x\sim\rho}\left[\left|\left|\sum_{i=1}^{B} g_{x_i}^2(\theta_t)\right|\right|_1\right] = B \cdot Tr\left(Cov_{x\sim\rho}(g_{x_i}(\theta_t))\right) + B||G(\theta_t)||_2^2 = B(\mu+1)||G(\theta_t)||_2^2. \tag{26}$$

Based on Assumption 3, we know that $g_{x_i}(\theta_t)$ and $g_{x_j}(\theta_t)$ are independent and identically distributed for $i \neq j$, i.e., $\mathbb{E}_{x\sim\rho}[g_{x_i}^T(\theta_t)g_{x_j}(\theta_t)] = (\mathbb{E}_{x\sim\rho}[g_x(\theta_t)])^T(\mathbb{E}_{x\sim\rho}[g_x(\theta_t)]), \forall i \neq j$. Therefore we can obtain that

$$\mathbb{E}_{x\sim\rho}\left[\left|\left|\sum_{i=1}^{B} g_{x_i}(\theta_t)\right|\right|_2^2\right] = \mathbb{E}_{x\sim\rho}\left[\left|\left|\sum_{i=1}^{B} g_{x_i}^2(\theta_t)\right|\right|_1\right] + \mathbb{E}_{x\sim\rho}\left[\sum_{1\leq i,j\leq B, i\neq j} g_{x_i}^T(\theta_t)g_{x_j}(\theta_t)\right] = \left(B\mu + B^2\right)||G(\theta_t)||_2^2. \tag{27}$$

By definition 2, we have

$$\mathbb{E}_x[GR_S] \approx \frac{\mathbb{E}_{x\sim\rho}\left[\left|\left|\sum_{i=1}^{B} g_{x_i}^2(\theta_t)\right|\right|_1\right]}{\mathbb{E}_{x\sim\rho}\left[\left|\left|\sum_{i=1}^{B} g_{x_i}(\theta_t)\right|\right|_2^2\right]} = \frac{1+\mu}{B+\mu}. \tag{28}$$

□

We next proceed to prove Theorem 2.

Theorem 2. *There exists a negative partial derivative relationship between the batch size B and the spatial gradient redundancy GR_S:*

$$\frac{\partial}{\partial B}\mathbb{E}[GR_S] < 0. \tag{29}$$

Proof. By Lemma 2, it's obvious that

$$\frac{\partial}{\partial B}\mathbb{E}[GR_S] = -\frac{1+\mu}{(B+\mu)^2} < 0. \tag{30}$$

□

D. The methods for practical GR_S measurements

Notice that it is impractical to compute the gradient for each sample individually. In data parallelism, we assume there are n devices, so the local batch size is set to B/n. Let the gradient aggregated within each device i be denoted as $\tilde{G}_{est,i}(\theta_t)$. Similar to GR_S, we define

$$\xi_k = \frac{||\sum_{i=1}^{k} \tilde{G}_{est,i}^2(\theta_t)||_1}{||\sum_{i=1}^{k} \tilde{G}_{est,i}(\theta_t)||_2^2} \tag{31}$$

for $k = 1, 2, \ldots, n$. Note that in data-parallel distributed training, we perform an all-reduce operation every iteration. This process involves collecting the gradients from all local batches and then reducing them. Through this process that we can easily compute $\sum_{i=1}^{k} \tilde{G}_{est,i}^2(\theta_t)$ and $\sum_{i=1}^{k} \tilde{G}_{est,i}(\theta_t)$ between devices when reducing the gradient, thus the calculation of ξ_k is feasible. According to Assumption 1 and Assumption 3, we can similarly obtain that

$$Tr(Cov_{x\sim\rho}(\tilde{G}_{est}(\theta))) = \frac{Tr(Cov_{x\sim\rho}(g_x(\theta)))}{B/n} = \frac{n}{B}\mu||G(\theta)||_2^2. \tag{32}$$

Similarly to Appendix C, we can easily obtain that

$$\mathbb{E}_{x\sim\rho}\left[\xi_k\right] \approx \frac{1+\frac{n\mu}{B}}{k+\frac{n\mu}{B}}. \tag{33}$$

Therefore, by transformation, we have

$$\frac{k-1}{1-\xi_k} - k = \frac{n}{B}\mu. \tag{34}$$

By performing linear fitting between $\left(\frac{k-1}{1-\xi_k} - k\right)$ and $\frac{n}{B}$ (for $k = 1, 2, \ldots, n$), the slope μ can be obtained when $n \geq 2$, from which $\mathbb{E}_x\left[GR_S\right]$ can be further derived from Lemma 2 which has been solidly validated. In the special case when $n = 1$ (only a single worker), we set the gradient accumulation steps to at least two, through which equivalent estimates across multiple devices can be obtained.

E. Theoretical estimate of the ideal value for GR_T and GR_S

We first estimate the theoretical ideal value for GR_S. Previous works [1] suggests that unlimitedly increasing the batch size does not yield proportional performance scaling. There exists a 'turning point' beyond which further batch size enlargement leads to rapidly diminishing scaling returns. The turning point is given as

$$\mathcal{B} = \frac{Tr(H \cdot Cov(\nabla L_i))}{g^T H g} \approx \frac{\mu\bar{\lambda}||g||_2^2}{\bar{\lambda}||g||_2^2} = \mu. \tag{35}$$

This suggests that batch size shouldn't exceed μ to avoid helpless and inefficient scale up. According to Lemma 2, we get the corresponding GR_S value with $B = \mathcal{B}$ is

$$\mathbb{E}_{x\sim\rho}[GR_S]_{B=\mathcal{B}} \approx \frac{1+\mu}{B+\mu} \approx 0.5. \tag{36}$$

Since $\frac{\partial}{\partial B}\mathbb{E}_{x\sim\rho}[GR_S] < 0$, this suggests that the ideal $\mathbb{E}_{x\sim\rho}[GR_S] \geq 0.5$.

On the other hand, to achieve efficient training, the batch size should not be too small. A typical lower-bound estimate is half of the turning point batch size (i.e., $\mathcal{B}/2$). Under this configuration, when applied to Lemma 2, we obtain $\mathbb{E}_{x\sim\rho}[GR_S] = \frac{1+\mu}{\mathcal{B}/2+\mu} \approx 0.65$. In summary, we theoretically estimate the GR_S range to be $[0.5, 0.65]$, which has been empirically validated by our preceding experiments.

Then we estimate the theoretical ideal value for GR_T. We take $B = \mu$ as the typical value. According to the gradient noise scale model [1], it gives the best learning rate with best efficiency by

$$\eta_{opt} = \frac{||g||_2^2}{g^T H g}\frac{1}{1+\frac{\mathcal{B}}{B}} \approx \frac{1}{2\bar{\lambda}}. \tag{37}$$

From Lemma 1, we obtain that

$$\mathbb{E}_{x\sim\rho}[GR_T]_{B=\mu,\beta=0.9,\eta=\frac{1}{2\bar{\lambda}}} = \left[\frac{(1+\frac{\mu}{B})^2(1-\bar{\lambda}\eta-\beta)^2}{(1-\beta)\left(2\bar{\lambda}\eta-(1-\beta)\left(1+\frac{\mu}{B}\right)\right)}\cdot\frac{\beta^{\frac{1}{1-\beta}}-\left(1-\frac{2\bar{\lambda}\eta}{1+\frac{\mu}{B}}\right)^{\frac{1}{1-\beta}}}{\left(\beta^{\frac{1}{1-\beta}}-\left(1-\bar{\lambda}\eta\right)^{\frac{1}{1-\beta}}\right)^2}\right]_{B=\mu,\beta=0.9,\eta=\frac{1}{2\bar{\lambda}}} \approx 20.$$

That suggests the ideal value of GR_T will be around 20.

References

[1] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018. 6