Supplementary Material to "Momentum Batch Normalization for Deep Learning with Small Batch Size"

Hongwei Yong[1,2], Jianqiang Huang[2], Deyu Meng[3,4], Xiansheng Hua[2], and Lei Zhang[1,2]

[1] Department of Computing, The Hong Kong Polytechnic University
{cshyong,cslzhang}@comp.polyu.edu.hk
[2] DAMO Academy, Alibaba Group
{jianqiang.jqh,huaxiansheng}@gmail.com
[3] Macau University of Science and Technology
[4] School of Mathematics and Statistics, Xi'an Jiaotong University
dymeng@mail.xjtu.edu.cn

In this supplementary file, we provide proofs of the theoretical results in the main paper, including Theorem 1 and Theorem 2.

A1. Proof of Theorem 1

Theorem 1 *Suppose samples x_i for $i = 1, 2, ..., m$ are i.i.d. with $E[x] = \mu$ and $Var[x] = \sigma^2$, ξ_μ and ξ_σ are defined in Eq.(2), we have:*

$$\lim_{m\to\infty} p(\xi_\mu) \to \mathcal{N}(0, \frac{1}{m}), \quad \lim_{m\to\infty} p(\xi_\sigma) \to \frac{1}{m}\chi^2(m-1).$$

Proof. From the classical central limit theorem, we have

$$\lim_{m\to\infty} p(\sum_{i=1}^{m} x_i) \to \mathcal{N}(m\mu, m\sigma^2),$$

that is

$$\lim_{m\to\infty} p(m\mu_B) \to \mathcal{N}(m\mu, m\sigma^2).$$

Therefore $\lim_{m\to\infty} p(\mu_B) \to \mathcal{N}(\mu, \frac{\sigma^2}{m})$, and $\xi_\mu = \frac{\mu - \mu_B}{\sigma}$ is a linear function of μ_B. Then according to the property of Gaussian distribution, we can obtain that

$$\lim_{m\to\infty} p(\xi_\mu) \to \mathcal{N}(0, \frac{1}{m}).$$

For χ^2 distribution, it has the following property:

$$\lim_{m\to\infty} \frac{\chi^2(m-1)}{m} = \lim_{m\to\infty} \frac{\chi^2(m-1)}{m-1} \to \mathcal{N}(1, \frac{2}{m}).$$

And we have $\lim_{m\to\infty} \mu_B = \mu$. Then for ξ_σ we can also use the central limit theorem to obtain:

$$\lim_{m\to\infty} p(\xi_\sigma) = \lim_{m\to\infty} p(\frac{1}{m\sigma^2}\sum_{i=1}^{m}(x_i-\mu_B)^2) = \lim_{m\to\infty} p(\sum_{i=1}^{m}(\frac{x_i-\mu}{\sigma})^2) \to \mathcal{N}(1, \frac{\kappa}{m}),$$

where κ is the kurtosis of x. When m is a very large number, both $\frac{\kappa}{m}$ and $\frac{2}{m}$ are close to zeros so that $\mathcal{N}(1, \frac{\kappa}{m}) \simeq \mathcal{N}(1, \frac{2}{m})$. Therefore, in this case the distribution of ξ_σ can be viewed as $\frac{1}{m}\chi^2(m-1)$.

The proof is completed. ■

A2. Proof of Theorem 2

Theorem 2: *If the infinite derivative of $l(x)$ exists for any x, given two random variables ξ_μ and ξ_σ (>0), then we have the Taylor expansion for $l(\frac{x+\xi_\mu}{\sqrt{\xi_\sigma}})$:*

$$E_{\xi_\mu,\xi_\sigma}[l(\frac{x+\xi_\mu}{\sqrt{\xi_\sigma}})] = l(x) + R^{add}(x) + R^{mul}(x) + R(x),\ R(x) = \sum_{n=1}^{\infty}\frac{E[{\xi_\mu}^n]}{n!}\frac{d^n R^{mul}(x)}{dx^n} \tag{1}$$

where $R^{add}(x)$ and $R^{mul}(x)$ are defined in Eq.(5) and (6), respectively.

Proof.

$$\begin{aligned}
E_{\xi_\mu,\xi_\sigma}[l(\frac{x+\xi_\mu}{\sqrt{\xi_\sigma}})] &= E_{\xi_\mu,\xi_\sigma}[l(\frac{x}{\sqrt{\xi_\sigma}} + \frac{\xi_\mu}{\sqrt{\xi_\sigma}})] \\
&= E_{\xi_\sigma}[l(\frac{x}{\sqrt{\xi_\sigma}})] + E_{\xi_\mu,\xi_\sigma}[\sum_{n=1}^{\infty}\frac{(\frac{\xi_\mu}{\sqrt{\xi_\sigma}})^n}{n!}\frac{d^n l(\frac{x}{\sqrt{\xi_\sigma}})}{d(\frac{x}{\sqrt{\xi_\sigma}})^n}] \\
&= l(x) + R^{mul}(x) + E_{\xi_\mu,\xi_\sigma}[\sum_{n=1}^{\infty}\frac{{\xi_\mu}^n}{n!}\frac{d^n l(\frac{x}{\sqrt{\xi_\sigma}})}{dx^n}] \\
&= l(x) + R^{mul}(x) + E_{\xi_\mu}[\sum_{n=1}^{\infty}\frac{{\xi_\mu}^n}{n!}\frac{d^n E_{\xi_\sigma}[l(\frac{x}{\sqrt{\xi_\sigma}})]}{dx^n}] \\
&= l(x) + R^{mul}(x) + E_{\xi_\mu}[\sum_{n=1}^{\infty}\frac{{\xi_\mu}^n}{n!}\frac{d^n (l(x) + R^{mul}(x))}{dx^n}] \\
&= l(x) + R^{mul}(x) + E_{\xi_\mu}[\sum_{n=1}^{\infty}\frac{{\xi_\mu}^n}{n!}\frac{d^n (l(x))}{dx^n}] + E_{\xi_\mu}[\sum_{n=1}^{\infty}\frac{{\xi_\mu}^n}{n!}\frac{d^n (R^{mul}(x))}{dx^n}] \\
&= l(x) + R^{mul}(x) + R^{add}(x) + R(x).
\end{aligned} \tag{2}$$

The proof is completed. ■