Hierarchically Robust Representation Learning
Supplementary

Qi Qian\textsuperscript{1} Juhua Hu\textsuperscript{2} Hao Li\textsuperscript{1}
\textsuperscript{1}Alibaba Group
\textsuperscript{2}School of Engineering and Technology
University of Washington, Tacoma, USA
\{qi.qian, lihao.lh\}@alibaba-inc.com jhuah@uw.edu

1. Proof of Theorem 1

\textbf{Proof.} Due to the smoothness, we have
\[ \ell(\hat{x}_i, y_i; \theta) \leq \ell(x_i, y_i; \theta) + \langle \nabla_x \ell, \hat{x}_i - x_i \rangle + \frac{L_x}{2} \| \hat{x}_i - x_i \|^2_F. \]
So
\[ \ell(\hat{x}_i, y_i; \theta) - \frac{\lambda_w}{2} \| \hat{x}_i - x_i \|^2_F \leq \ell(x_i, y_i; \theta) + \langle \nabla_x \ell, \hat{x}_i - x_i \rangle - \frac{\lambda_w - L_x}{2} \| \hat{x}_i - x_i \|^2_F. \]
When \( \lambda_w \) is sufficiently large as \( \lambda_w > L_x \), R.H.S. is bounded and
\[ \ell(\hat{x}_i, y_i; \theta) - \frac{\lambda_w}{2} \| \hat{x}_i - x_i \|^2_F \leq \ell(x_i, y_i; \theta) + \frac{1}{2(\lambda_w - L_x)} \| \nabla_x \ell \|^2_F. \]
Since \( \nabla_x \ell(\cdot) \) is \( L_{\theta} \)-Lipschitz continuous, we have
\[ \| \nabla_x \ell(x; \theta) \|^2_F \leq 2 \| \nabla_x \ell(x; \theta) - \nabla_x \ell(x; 0) \|^2_F + 2 \| \nabla_x \ell(x; 0) \|^2_F \leq 2L_{\ell, \theta}^2 \| \theta \|^2_F + 2 \| \nabla_x \ell(x; 0) \|^2_F. \]
Note that \( \| \nabla_x \ell(x; 0) \|^2_F = 0 \) in many convolutional neural networks. The bound can be improved and the original subproblem can be bounded as
\[ \max_{x_i \in X} \ell(x_i, y_i; \theta) - \frac{\lambda_w}{2} \| x_i - x_i \|^2_F \leq \ell(x_i, y_i; \theta) + \frac{\gamma}{2} \| \theta \|^2_F, \]
where \( \gamma = \frac{L_{\ell, \theta}^2}{\lambda_w - L_x}. \)

2. Proof of Theorem 2

\textbf{Proof.} We consider the augmented examples as \( \tilde{x}_i = x_i + \tau z_i \)
According to the smoothness, we have
\[ \ell(\tilde{x}_i, y_i; \theta) - \frac{\lambda_w}{2} \| \tilde{x}_i - x_i \|^2 \leq \ell(\tilde{x}_i, y_i; \theta) + \langle \nabla_{\tilde{x}} \ell, \tilde{x}_i - x_i \rangle. \]
\[ \frac{L_x}{2} \| \tilde{x}_i - x_i \|^2 - \frac{\lambda_w}{2} \| \tilde{x}_i - x_i \|^2 \leq \ell(\tilde{x}_i, y_i; \theta) + \gamma \| \theta \|^2_F \]
\[ \leq \ell(\tilde{x}_i, y_i; \theta) + \frac{\tau^2 L_x \| z_i \|^2_F}{2(\lambda_w - L_x)} - \tau \langle \nabla_{\tilde{x}} \ell, z_i \rangle. \]
\[ \leq \ell(\tilde{x}_i, y_i; \theta) + \frac{\tau^2 L_x \| z_i \|^2_F}{2(\lambda_w - L_x)} - \tau \langle \nabla_{\tilde{x}} \ell, z_i \rangle. \]
The last equation is from setting \( \tau \) to optimize as
\[ \tau = \frac{\langle \nabla_{\tilde{x}} \ell, z_i \rangle}{3L_x \| z_i \|^2_F}. \]
3. Proof of Theorem 3

Proof. For an arbitrary distribution $q$, we have

$$
E[||q_{t+1} - q||_2^2] = E[||P_\Delta(q_t + \eta_t g_t) - q||_2^2] \\
\leq E[||q_t + \eta_t g_t - q||_2^2] \\
= E[||q_t - q||_2^2 + 2\eta_t (q_t - q)^T g_t + \eta_t^2 ||g_t||_2^2] \\
\leq E[||q_t - q||_2^2 + \eta_t^2 \mu^2] \\
+ 2\eta_t (\mathcal{L}(q_t, \theta_t) - \mathcal{L}(q_t, \theta_t) - \frac{\lambda}{2} ||q_t - q||_2^2)
$$

By setting $c = \frac{\mu}{\lambda} \sqrt{\frac{\log(s)}{2s}}$ and $q$ to be optimum, we have

$$
\max_{q^* \in \Delta} \sum_t E[\mathcal{L}(q^*, \theta_t) - \mathcal{L}(q_t, \theta_t)] \leq \mu \sqrt{2s \log(s)} - s \lambda \\
- \frac{\mu^2 \log(s)}{2\lambda} + \frac{\mu^2}{2\lambda} (\log(T) + 1) \\
= \frac{\mu^2}{2\lambda} (\log(T) + 1) - (\mu \sqrt{\frac{\log(s)}{2s}} - \sqrt{s \lambda})^2
$$

The last inequality is from the fact that the objective is $\lambda$-strongly concave in $q$ and the observed gradient is unbiased. Therefore, we have

$$
E[\mathcal{L}(q_t, \theta_t) - \mathcal{L}(q_t, \theta_t)] \leq \frac{E[||q_t - q||_2^2] - E[||q_{t+1} - q||_2^2]}{2\eta_t} \\
- \frac{\lambda}{2} ||q_t - q||_2^2 + \frac{\eta_t}{2} \mu^2
$$

When $\eta_t = \frac{1}{\lambda t}$, we have

$$
E[\mathcal{L}(q_t, \theta_t) - \mathcal{L}(q_t, \theta_t)] \leq \frac{\lambda t}{2} (E[||q_t - q||_2^2] - E[||q_{t+1} - q||_2^2]) \\
- \frac{\lambda}{2} ||q_t - q||_2^2 + \frac{1}{2\lambda t} \mu^2
$$

When $\eta_t = \frac{1}{\lambda tc}$, we have

$$
E[\mathcal{L}(q_t, \theta_t) - \mathcal{L}(q_t, \theta_t)] \leq \frac{\lambda tc}{2} (E[||q_t - q||_2^2] - E[||q_{t+1} - q||_2^2]) \\
- \frac{\lambda}{2} ||q_t - q||_2^2 + \frac{1}{2\lambda tc} \mu^2
$$

We assume that $\eta_t = \frac{1}{\lambda t}$ and $c > 1$ for the first $s$ iterations and then $\eta_t = \frac{1}{\lambda t}$. So we have

$$
\sum_t^T E[\mathcal{L}(q, \theta_t) - \mathcal{L}(q_t, \theta_t)] = \sum_{t=1}^s E[\mathcal{L}(q, \theta_t) - \mathcal{L}(q_t, \theta_t)] \\
+ \sum_{t=s+1}^T E[\mathcal{L}(q, \theta_t) - \mathcal{L}(q_t, \theta_t)] \\
\leq \sum_{t=1}^s \left( \frac{c\lambda}{2} - \frac{\lambda}{2} \right) E[||q_t - q||_2^2] + \frac{1}{2\lambda tc} \mu^2 \\
+ \sum_{t=s+1}^T \frac{1}{2\lambda t} \mu^2 \\
\leq s(\lambda(c - 1) + \frac{\mu^2}{2\lambda} \log(s)(\frac{1}{c} - 1) + \frac{\mu^2}{2\lambda} (\log(T) + 1))
$$