Hierarchically Robust Representation Learning Supplementary

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

1. Proof of Theorem 1

Proof. Due to the smoothness, we have

$$\ell(\hat{\mathbf{x}}_i, y_i;\theta) \leq \ell(\mathbf{x}_i, y_i;\theta) + \langle \nabla_{\mathbf{x}_i}\ell, \hat{\mathbf{x}}_i - \mathbf{x}_i\rangle + \frac{L_{\mathbf{x}}}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_F^2$$

So

$$\ell(\hat{\mathbf{x}}_i, y_i;\theta) - \frac{\lambda_w}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_F^2 \leq \ell(\mathbf{x}_i, y_i;\theta) + \langle \nabla_{\mathbf{x}_i}\ell, \hat{\mathbf{x}}_i - \mathbf{x}_i\rangle - \frac{\lambda_w - L_{\mathbf{x}}}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_F^2$$

When λ_w is sufficiently large as $\lambda_w > L_{\mathbf{x}}$, R.H.S. is bounded and

$$\ell(\hat{\mathbf{x}}_i, y_i;\theta) - \frac{\lambda_w}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_F^2 \leq \ell(\mathbf{x}_i, y_i;\theta) + \frac{1}{2(\lambda_w - L_{\mathbf{x}})}\|\nabla_{\mathbf{x}_i}\ell\|_F^2$$

Since $\nabla_{\mathbf{x}}\ell(\cdot)$ is L_θ-Lipschitz continuous, we have

$$\begin{aligned}\|\nabla_{\mathbf{x}}\ell(\mathbf{x};\theta)\|_F^2 &\leq 2\|\nabla_{\mathbf{x}}\ell(\mathbf{x};\theta) - \nabla_{\mathbf{x}}\ell(\mathbf{x};\mathbf{0})\|_F^2 \\ &\quad + 2\|\nabla_{\mathbf{x}}\ell(\mathbf{x};\mathbf{0})\|_F^2 \\ &\leq 2L_\theta^2\|\theta\|_F^2 + 2\|\nabla_{\mathbf{x}}\ell(\mathbf{x};\mathbf{0})\|_F^2\end{aligned}$$

Note that $\|\nabla_{\mathbf{x}}\ell(\mathbf{x};\mathbf{0})\|_F = 0$ in many convolutional neural networks. The bound can be improved and the original subproblem can be bounded as

$$\max_{\hat{\mathbf{x}}_i \in \mathbf{X}} \ell(\hat{\mathbf{x}}_i, y_i;\theta) - \frac{\lambda_w}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_F^2 \leq \ell(\mathbf{x}_i, y_i;\theta) + \frac{\gamma}{2}\|\theta\|_F^2$$

where $\gamma = \frac{L_\theta^2}{\lambda_w - L_{\mathbf{x}}}$. □

2. Proof of Theorem 2

Proof. We consider the augmented examples as

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \tau\mathbf{z}_i$$

According to the smoothness, we have

$$\begin{aligned}&\ell(\hat{\mathbf{x}}_i, y_i;\theta) - \frac{\lambda_w}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 \leq \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \langle \nabla_{\tilde{\mathbf{x}}_i}\ell, \hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_i\rangle \\ &+ \frac{L_{\mathbf{x}}}{2}\|\hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_i\| - \frac{\lambda_w}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 \\ &= \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \langle \nabla_{\tilde{\mathbf{x}}_i}\ell - \tau L_{\mathbf{x}}\mathbf{z}_i, \hat{\mathbf{x}}_i - \mathbf{x}_i\rangle - \tau\langle \nabla_{\tilde{\mathbf{x}}_i}\ell, \mathbf{z}_i\rangle \\ &+ \frac{L_{\mathbf{x}}\tau^2}{2}\|\mathbf{z}_i\|^2 - \frac{\lambda_w - L_{\mathbf{x}}}{2}\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 \\ &\leq \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \frac{\|\nabla_{\tilde{\mathbf{x}}_i}\ell - \tau L_{\mathbf{x}}\mathbf{z}_i\|_F^2}{2(\lambda_w - L_{\mathbf{x}})} - \tau\langle \nabla_{\tilde{\mathbf{x}}_i}\ell, \mathbf{z}_i\rangle \\ &+ \frac{L_{\mathbf{x}}\tau^2}{2}\|\mathbf{z}_i\|_F^2 \\ &= \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \frac{\|\nabla_{\tilde{\mathbf{x}}_i}\ell\|_F^2}{2(\lambda_w - L_{\mathbf{x}})} \\ &+ \frac{\lambda_w}{\lambda_w - L_{\mathbf{x}}}(\frac{\tau^2 L_{\mathbf{x}}\|\mathbf{z}_i\|_F^2}{2} - \tau\langle \nabla_{\tilde{\mathbf{x}}_i}\ell, \mathbf{z}_i\rangle) \\ &\leq \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \frac{\gamma}{2}\|\theta\|_F^2 \\ &+ \frac{\lambda_w}{\lambda_w - L_{\mathbf{x}}}(\frac{\tau^2 L_{\mathbf{x}}\|\mathbf{z}_i\|_F^2}{2} - \tau\langle \nabla_{\tilde{\mathbf{x}}_i}\ell - \nabla_{\mathbf{x}_i}\ell, \mathbf{z}_i\rangle - \tau\langle \nabla_{\mathbf{x}_i}\ell, \mathbf{z}_i\rangle) \\ &\leq \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \frac{\gamma}{2}\|\theta\|_F^2 \\ &+ \frac{\lambda_w}{\lambda_w - L_{\mathbf{x}}}(\frac{\tau^2 L_{\mathbf{x}}\|\mathbf{z}_i\|_F^2}{2} + \tau\|\nabla_{\tilde{\mathbf{x}}_i}\ell - \nabla_{\mathbf{x}_i}\ell\|\|\mathbf{z}_i\| - \tau\langle \nabla_{\mathbf{x}_i}\ell, \mathbf{z}_i\rangle) \\ &\leq \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \frac{\gamma}{2}\|\theta\|_F^2 \\ &+ \frac{\lambda_w}{\lambda_w - L_{\mathbf{x}}}(\frac{3\tau^2 L_{\mathbf{x}}\|\mathbf{z}_i\|_F^2}{2} - \tau\langle \nabla_{\mathbf{x}_i}\ell, \mathbf{z}_i\rangle) \\ &= \ell(\tilde{\mathbf{x}}_i, y_i;\theta) + \frac{\gamma}{2}\|\theta\|_F^2 - \frac{\lambda_w}{\lambda_w - L_{\mathbf{x}}}\frac{\langle \nabla_{\mathbf{x}_i}\ell, \mathbf{z}_i\rangle^2}{6L_{\mathbf{x}}\|\mathbf{z}_i\|_F^2}\end{aligned}$$

The last equation is from setting τ to optimum as

$$\tau = \frac{\langle \nabla_{\mathbf{x}_i}\ell, \mathbf{z}_i\rangle}{3L_{\mathbf{x}}\|\mathbf{z}_i\|_F^2}$$

□

3. Proof of Theorem 3

Proof. For an arbitrary distribution $\mathbf{q}$, we have

$$
\begin{aligned}
&E[\|\mathbf{q}_{t+1}-\mathbf{q}\|_2^2] = E[\|\mathcal{P}_\Delta(\mathbf{q}_t+\eta_t g_t)-\mathbf{q}\|_2^2] \\
&\le E[\|\mathbf{q}_t+\eta_t g_t-\mathbf{q}\|_2^2] \\
&= E[\|\mathbf{q}_t-\mathbf{q}\|_2^2 + 2\eta_t(\mathbf{q}_t-\mathbf{q})^\top g_t + \eta_t^2\|g_t\|_2^2] \\
&\le E[\|\mathbf{q}_t-\mathbf{q}\|_2^2 + \eta_t^2\mu^2 \\
&+ 2\eta_t(\mathcal{L}(\mathbf{q}_t,\theta_t)-\mathcal{L}(\mathbf{q},\theta_t)-\frac{\lambda}{2}\|\mathbf{q}_t-\mathbf{q}\|_2^2)]
\end{aligned}
$$

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

$$
\begin{aligned}
&E[\mathcal{L}(\mathbf{q},\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] \le \frac{E[\|\mathbf{q}_t-\mathbf{q}\|_2^2]-E[\|\mathbf{q}_{t+1}-\mathbf{q}\|_2^2]}{2\eta_t} \\
&-\frac{\lambda}{2}\|\mathbf{q}_t-\mathbf{q}\|_2^2+\frac{\eta_t}{2}\mu^2
\end{aligned}
$$

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

$$
\begin{aligned}
&E[\mathcal{L}(\mathbf{q},\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] \le \frac{\lambda t}{2}(E[\|\mathbf{q}_t-\mathbf{q}\|_2^2]-E[\|\mathbf{q}_{t+1}-\mathbf{q}\|_2^2]) \\
&-\frac{\lambda}{2}\|\mathbf{q}_t-\mathbf{q}\|_2^2+\frac{1}{2\lambda t}\mu^2
\end{aligned}
$$

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

$$
\begin{aligned}
&E[\mathcal{L}(\mathbf{q},\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] \le \frac{\lambda tc}{2}(E[\|\mathbf{q}_t-\mathbf{q}\|_2^2]-E[\|\mathbf{q}_{t+1}-\mathbf{q}\|_2^2]) \\
&-\frac{\lambda}{2}\|\mathbf{q}_t-\mathbf{q}\|_2^2+\frac{1}{2\lambda tc}\mu^2
\end{aligned}
$$

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

$$
\begin{aligned}
&\sum_t^T E[\mathcal{L}(\mathbf{q},\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] = \sum_{t=1}^s E[\mathcal{L}(\mathbf{q},\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] \\
&+\sum_{t=s+1}^T E[\mathcal{L}(\mathbf{q},\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] \\
&\le \sum_{t=1}^s\left((\frac{c\lambda}{2}-\frac{\lambda}{2})E[\|\mathbf{q}_t-\mathbf{q}\|_2^2]+\frac{1}{2\lambda tc}\mu^2\right)+\sum_{t=s+1}^T\frac{1}{2\lambda t}\mu^2 \\
&\le s\lambda(c-1)+\frac{\mu^2}{2\lambda}\log(s)(\frac{1}{c}-1)+\frac{\mu^2}{2\lambda}(\log(T)+1)
\end{aligned}
$$

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

$$
\begin{aligned}
&\max_{\mathbf{q}^*\in\Delta}\sum_t^T E[\mathcal{L}(\mathbf{q}^*,\theta_t)-\mathcal{L}(\mathbf{q}_t,\theta_t)] \le \mu\sqrt{2s\log(s)}-s\lambda \\
&-\frac{\mu^2\log(s)}{2\lambda}+\frac{\mu^2}{2\lambda}(\log(T)+1) \\
&=\frac{\mu^2}{2\lambda}(\log(T)+1)-(\mu\sqrt{\frac{\log(s)}{2\lambda}}-\sqrt{s\lambda})^2
\end{aligned}
$$

□