Tight Sample Complexity of Learning One-hidden-layer Convolutional Neural Networks

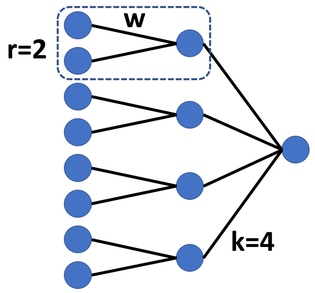
Yuan Cao and **Quanquan Gu**
Department of Computer Science, University of California, Los Angeles

NEURAL INFORMATION PROCESSING SYSTEMS

One-hidden-layer Convolutional Neural Networks

- **CNN with no filter overlapping**

$$y = \sum_{j=1}^{k} v_j \sigma(\mathbf{w}^\top \mathbf{P}_j \mathbf{x}_i),$$



- r: filter size
- k: number of neurons
- $\mathbf{w}$: convolution filter
- $\mathbf{v}$: second layer weight
- $\sigma(\cdot)$: activation function

Questions We Aim to Answer

Can convergence guarantee be established for general activation functions?

What is the sample complexity for learning CNNs with non-overlapping filters?

Approximate gradient descent

Algorithm Approximate Gradient Descent for Non-overlapping CNN

Require: Training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, number of iterations T, step size α, initialization $\mathbf{w}^0 \in S^{r-1}$, $\mathbf{v}^0$.

for $t = 0, 1, 2, \ldots, T-1$ **do**

$\mathbf{g}_w^t = -\frac{1}{n}\sum_{i=1}^n \big[y_i - \sum_{j=1}^k v_j^t \sigma(\mathbf{w}^{t\top}\mathbf{P}_j\mathbf{x}_i)\big] \cdot \sum_{j'=1}^k \xi^{-1} v_{j'}^t \mathbf{P}_{j'}\mathbf{x}_i$

$\mathbf{g}_v^t = -\frac{1}{n}\boldsymbol{\Sigma}^\top(\mathbf{w}^t)[\mathbf{y} - \boldsymbol{\Sigma}(\mathbf{w}^t)\mathbf{v}^t]$

$\mathbf{u}^{t+1} = \mathbf{w}^t - \alpha \mathbf{g}_w^t$, $\mathbf{w}^{t+1} = \mathbf{u}^{t+1}/\|\mathbf{u}^{t+1}\|_2$, $\mathbf{v}^{t+1} = \mathbf{v}^t - \alpha \mathbf{g}_v^t$

end for

Ensure: $\mathbf{w}^T$, $\mathbf{v}^T$

$\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\Sigma(\mathbf{w}) = [\sigma(\mathbf{w}^\top \mathbf{P}_j \mathbf{x}_i)]_{n\times k}$, $\xi = \mathbb{E}_{z\sim N(0,1)}[\sigma(z)z]$.

Assumptions

- **Standard Gaussian inputs** The input data are generated from standard Gaussian distribution: $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$.
- **General activation functions** $\sigma(\cdot)$ is any 1-Lipschitz continuous and non-trival increasing function.
- **Teacher network** $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are generated from a teacher network with Gaussian noise $\{\epsilon_i\}$:

$$y_i = \sum_{j=1}^{k} v_j^* \sigma(\mathbf{w}^{*\top}\mathbf{P}_j\mathbf{x}_i) + \epsilon_i.$$

Linear Convergence Up to Statistical Precision

Let $\delta \in (0,1)$, $\gamma_1 = (1+\alpha\rho)^{-1/2}$, $\gamma_2 = \sqrt{1-\alpha\Delta+4\alpha^2\Delta^2}$, and $\gamma_3 = 1-\alpha(\Delta+\kappa^2 k)$. Suppose that $(\mathbf{w}^0, \mathbf{v}^0)$ satisfies

$$\mathbf{w}^{*\top}\mathbf{w}^0 > 0,\ \mathbf{v}^{*\top}\mathbf{v}^0 > 0,\ \kappa^2(\mathbf{1}^\top\mathbf{v}^*)\mathbf{1}^\top(\mathbf{v}^0-\mathbf{v}^*) \le \rho,$$

the step size $\alpha \le Q_1 k^{-1}$, and the sample size $n \ge Q_2\sqrt{r+k}$. Then with probability at least $1-\delta$,

$$\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \gamma_1^t\|\mathbf{w}^0-\mathbf{w}^*\|_2 + 8\rho^{-1}\gamma_1^{-2}\eta_w,$$
$$\|\mathbf{v}^t - \mathbf{v}^*\|_2 \le R_1 t^{3/2}\overline{\gamma}^t + R_2(1+|\kappa|\sqrt{k})(\eta_w+\eta_v),$$

for all $t = 0, \ldots, T$, where

$$\eta_w = R_3\sqrt{\frac{(r+k)\log(120nk/\delta)}{n}},$$
$$\eta_v = R_4(1+\kappa k)\cdot\sqrt{\frac{(r+k)\log(120nk/\delta)}{n}}.$$

ρ, Q_1, Q_2, R_1, R_2, R_3, R_4 are constants that only depend on $\sigma(\cdot)$, $(\mathbf{w}^*, \mathbf{v}^*)$ and $(\mathbf{w}^0, \mathbf{v}^0)$. $\kappa = \mathbb{E}_{z\sim N(0,1)}[\sigma(z)]$, $\Delta = \mathrm{Var}_{z\sim N(0,1)}[\sigma(z)]$, and $\overline{\gamma} = \gamma_1 \vee \gamma_2 \vee \gamma_3$.

Key Ingredients for the Proof

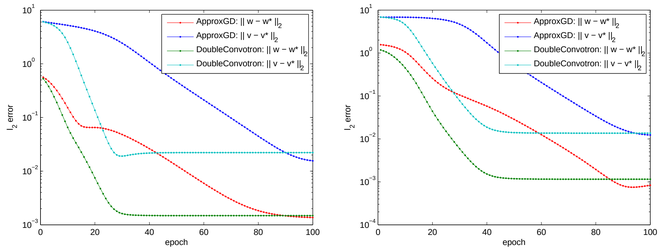
- **Gaussian comparison inequality**
  There exists an increasing function $\psi(\tau)$ such that
  $$\psi(\mathbf{w}^\top\mathbf{w}') = \mathrm{Cov}_{\mathbf{z}\sim N(\mathbf{0},\mathbf{I})}[\sigma(\mathbf{w}^\top\mathbf{z}), \sigma(\mathbf{w}'^\top\mathbf{z})],$$
  and $\Delta \ge \psi(\tau) > 0$ for all $\tau > 0$.
- **Uniform concentration inequalities for approximate gradients**
  There exist $\mathcal{W}_0, \mathcal{V}_0$ such that with high probability,
  $$\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}_0\times\mathcal{V}_0} \|\mathbf{g}_w(\mathbf{w},\mathbf{v}) - \overline{\mathbf{g}}_w(\mathbf{w},\mathbf{v})\|_2 \le \eta_w,$$
  $$\sup_{(\mathbf{w},\mathbf{v})\in\mathcal{W}_0\times\mathcal{V}_0} \|\mathbf{g}_v(\mathbf{w},\mathbf{v}) - \overline{\mathbf{g}}_v(\mathbf{w},\mathbf{v})\|_2 \le \eta_v,$$
  $$\overline{\mathbf{g}}_w(\mathbf{w},\mathbf{v}) = \|\mathbf{v}\|_2^2\mathbf{w} - (\mathbf{v}^{*\top}\mathbf{v})\mathbf{w}^*,$$
  $$\overline{\mathbf{g}}_v(\mathbf{w},\mathbf{v}) = (\Delta\mathbf{I} + \kappa^2\mathbf{1}\mathbf{1}^\top)\mathbf{v} - [\psi(\mathbf{w}^\top\mathbf{w}^*)\mathbf{I} + \kappa^2\mathbf{1}\mathbf{1}^\top]\mathbf{v}^*.$$
- **Convergence analysis based on $\overline{\mathbf{g}}_w, \overline{\mathbf{g}}_v$**
  There exist $\mathcal{W} \subseteq \mathcal{W}_0$ and $\mathcal{V} \subseteq \mathcal{V}_0$, such that
  - If $(\mathbf{w}^t, \mathbf{v}^t) \in \mathcal{W}\times\mathcal{V}$, then $(\mathbf{w}^{t+1}, \mathbf{v}^{t+1}) \in \mathcal{W}\times\mathcal{V}$.
  - Bounds of $\|\mathbf{w}^s - \mathbf{w}^*\|_2$ and $\|\mathbf{v}^s - \mathbf{v}^*\|_2$, $s \in [t]$ imply bounds for $\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2$ and $\|\mathbf{v}^{t+1} - \mathbf{v}^*\|_2$.

Comparison with Related Results

	Conv. rate	Sample comp.	Act. fun.	Data input	Overlap	Sec. layer
Du et al. [13]	linear	-	ReLU	Gaussian	no	yes
Du et al. [14]	-	$\widetilde{O}((k+r)\cdot\epsilon^{-2})$	Linear	sub-Gaussian	yes	-
Convotron [18]	(sub)linear	$\widetilde{O}(k^2 r\cdot\epsilon^{-2})$	(leaky) ReLU	symmetric	yes	no
D. Convotron [10]	sublinear	$\widetilde{O}(\mathrm{poly}(k,r,\epsilon^{-1}))$	(leaky) ReLU	symmetric	yes	yes
This paper	linear	$\widetilde{O}((k+r)\cdot\epsilon^{-2})$	general	Gaussian	no	yes

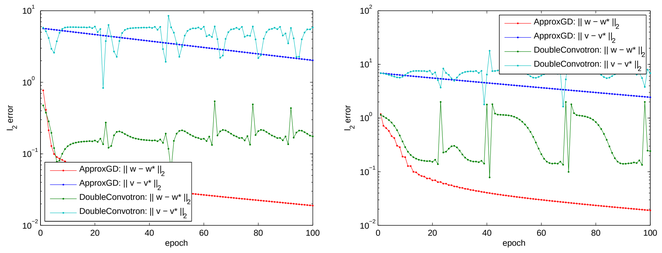
Numerical Experiments

- **Experiment setup**
  - Number of iterations $T = 100$, sample size $n = 1000$.
  - ReLU: $\alpha = 0.04$, sigmoid: $\alpha = 0.25$, tanh: $\alpha = 0.1$.
  - $\mathbf{w}^{(0)} \sim \mathrm{Unif}(S^{r-1})$. $\mathbf{v}^{(0)}$: in $\mathcal{B}(0, k^{-1/2}|\mathbf{1}^\top\mathbf{v}^*|/2)$. We present the best result among $\{(\pm\mathbf{w}^{(0)}, \pm\mathbf{v}^{(0)})\}$.
- **ReLU**



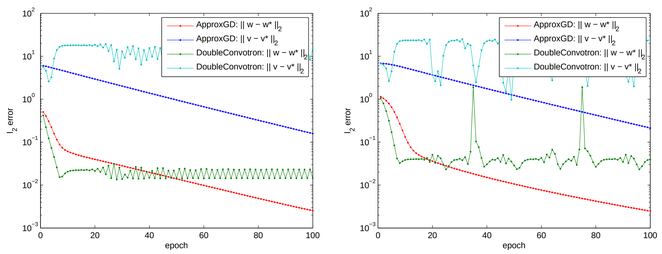
(a) $k = 15$, $r = 5$ (b) $k = 30$, $r = 9$

- **Sigmoid**



(c) $k = 15$, $r = 5$ (d) $k = 30$, $r = 9$

- **Hyper tangent**



(e) $k = 15$, $r = 5$ (f) $k = 30$, $r = 9$