# A Technical Proof

We prove Theorem 4 and Theorem 9 in this section.

## A.1 Proof of Theorem 4

We follows Dai et al. (2014) to decompose the error into two terms,

$$|f_t(\mathbf{x}) - f^*(\mathbf{x})|^2 \leq 2|f_t(\mathbf{x}) - h_t(\mathbf{x})|^2 + 2\kappa\|h_t - f^*\|_{\mathcal{H}}^2,$$

where

$$h_t(\cdot) = \mathbb{E}_{\boldsymbol{\omega}}[f_t(\cdot)] = \mathbb{E}_{\boldsymbol{\omega}}[\sum_{i=1}^{t} a_t^i \zeta_{\boldsymbol{\omega}_i}(\cdot)] = \sum_{i=1}^{t} a_t^i \xi_{\boldsymbol{\omega}_i}(\cdot).$$

**Lemma 1.** *(Dai et al., 2014) Assume $\ell'(u, y)$ is L-Lipschitz continuous in terms of $u \in \mathbb{R}$. Let $f_*$ be the optimal solution to our target problem. Then if we set $\gamma_t = \frac{\theta}{t}$ with $\theta$ such that $\theta\nu \in (1, 2) \cup \mathbb{Z}_+$, and $\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left(|f_{t+1}(\mathbf{x}) - h_{t+1}(\mathbf{x})|^2\right) \leq \frac{C^2}{t}$ then*

$$\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\|h_{t+1} - f_*\|_{\mathcal{H}}^2\right] \leqslant \frac{S^2}{t},$$

*where*

$$S = \max\left\{\|f_*\|_{\mathcal{H}}, \frac{Q+\sqrt{Q^2+Z(1+\theta\nu)^2\theta^2\kappa M^2}}{Z}\right\},$$

$Z = 2\lambda\theta - 1$, $Q = \sqrt{2}\kappa^{1/2}LC\theta$.

With Lemma 1, the remaining task is to bound $\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left(|f_{t+1}(\mathbf{x}) - h_{t+1}(\mathbf{x})|^2\right)$. We define the following terms to simplify the notations.

**Definition 2.**

- $\mathcal{G} = \{x | (x-1) \bmod (G+U) < G \text{ and } 1 \leq x \leq t\}$.
- $\mathcal{G}_k = \{x | x \in \mathcal{G} \text{ and } \lceil x/(G+U)\rceil = k\}$.
- $\mathcal{U}_k = \{x | x \notin \mathcal{G}, \lceil x/(G+U)\rceil = k \text{ and } 1 \leq x \leq t\}$.
- $\delta_i(\mathbf{x}) = \zeta_t(\mathbf{x}) - \xi_t(\mathbf{x})$
- $V_i(\mathbf{x}) = a_t^i\delta_i(\mathbf{x}) \leq c_i = |a_t^i|u$, *where* $u = 2M(\phi + \kappa)$.

By definition, we have $\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left(|f_{t+1}(\mathbf{x}) - h_{t+1}(\mathbf{x})|^2\right) = \mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{i=1}^t V_i(\mathbf{x})\right)^2\right]$, then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{i=1}^t V_i(\mathbf{x})\right)^2\right] \\
=\ &\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{\mathcal{G}} V_i(\mathbf{x}) + \sum_{k=1}^{\lceil t/(G+U)\rceil}\sum_{i\in\mathcal{U}_k} V_i(\mathbf{x})\right)^2\right] \\
\leq\ &\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{\mathcal{G}} V_i(\mathbf{x})\right)^2 + \sum_{k=1}^{\lceil t/(G+U)\rceil}\left(\sum_{i\in\mathcal{U}_k} V_i(\mathbf{x})\right)^2\right. \\
&+2\sum_{k=1}^{\lceil t/(G+U)\rceil}\left(\sum_{i\in\mathcal{U}_k} V_i(\mathbf{x})\right)\left(\sum_{\mathcal{G}} V_i(\mathbf{x})\right) \\
&\left.\sum_{p=1}^{\lceil \frac{t}{G+U}\rceil - 1}\sum_{q=p+1}^{\lceil \frac{t}{G+U}\rceil}\left(\sum_{i\in\mathcal{U}_p} V_i(\mathbf{x})\right)\left(\sum_{i\in\mathcal{U}_q} V_i(\mathbf{x})\right)\right].
\end{aligned}$$

Applying Lemma 3, 5 and 6 to bound each term yields

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{i=1}^t V_i(\mathbf{x})\right)^2\right] \\
\leq\ &\sum_{i=1}^t c_i^2 + \frac{\theta^2u^2U(U-1)}{Bt^2}\left(\sum_{k=1}^{\lceil \frac{t}{G+U}\rceil}\frac{1}{k^2}\right) + \\
&\frac{\theta^2u^2U}{(G+U)t} + \frac{2\theta^2u^2U^2}{Bt^2}\sum_{p=1}^{\lceil \frac{t}{G+U}\rceil - 1}\sum_{q=p+1}^{\lceil \frac{t}{G+U}\rceil}\frac{1}{q} \\
\leq\ &\left(1 + \frac{2U}{G+U} + \frac{2U^2}{B(G+U)}\right)\frac{\theta^2u^2}{t} + \frac{2U(U-1)}{Gt}\frac{\theta^2u^2}{t},
\end{aligned}$$

where the second inequality is by Lemma 7.

**Lemma 3.** *If $i \in \mathcal{U}_k$, then*

$$\mathbb{E}\left[\left(\sum_{i\in\mathcal{U}_k} V_i(\mathbf{x})\right)^2\right] \leq \sum_{i\in\mathcal{U}_k} c_i^2 + \frac{U(U-1)\theta^2u^2}{Bk^2t^2}.$$

*Proof.* By expanding the quadratic term in the expectation,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i\in\mathcal{U}_k} V_i(\mathbf{x})\right)^2\right] \\
\leq\ &\mathbb{E}\left(\sum_{i\in\mathcal{U}_k} V_i^2(\mathbf{x}) + \sum_{i,j\in\mathcal{U}_k, i\neq j} V_i(\mathbf{x})V_j(\mathbf{x})\right) \\
\leq\ &\sum_{i\in\mathcal{U}_k} c_i^2 + \sum_{i,j\in\mathcal{U}_k, i\neq j}\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\frac{a_t^i}{Gk}\sum_{ii\in\mathcal{G}_k}\delta_{ii}(\mathbf{x})\right)\right. \\
&\left.\left(\frac{a_t^j}{Gk}\sum_{ii\in\mathcal{G}_k}\delta_{ii}(\mathbf{x})\right)\right] \\
\leq\ &\sum_{i\in\mathcal{U}_k} c_i^2 + \frac{U(U-1)\theta^2u^2}{Bk^2t^2}.
\end{aligned}$$

The second inequality is by Definition 2 and the third inequality is by Lemma 8 and Lemma 4. □

**Lemma 4.** *(Dai et al., 2014) Suppose $\gamma_i = \frac{\theta}{i}$, where $1 \le i \le t$ and $\theta\nu \in (1,2) \cup \mathbb{Z}_+$, then $a_t^i \le \frac{\theta}{t}$.*

**Lemma 5.** *If $i \in \mathcal{U}_k$, then*

$$\mathbb{E}\left(V_i(\mathbf{x})\sum_{j\in\mathcal{G}} V_j(\mathbf{x})\right) \le \frac{\theta^2 u^2}{t^2}.$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}\left(V_i(\mathbf{x})\sum_{j\in\mathcal{G}} V_j(\mathbf{x})\right)\\
=\ & \mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\frac{1}{Bk}\left(\sum_{ii\in\mathcal{G}_k} a_t^{ii}\delta_{ii}(\mathbf{x})\right)\left(\sum_{jj\in\mathcal{G}} a_t^{jj}\delta_{jj}(\mathbf{x})\right)\right]\\
\le\ & \frac{\theta^2 u^2}{t^2}.
\end{aligned}$$

The inequality is by Lemma 8 and Lemma 4. □

**Lemma 6.** *If $p < q$, then*

$$\mathbb{E}\left[\left(\sum_{i\in\mathcal{U}_p} V_i(\mathbf{x})\right)\left(\sum_{j\in\mathcal{U}_q} V_j(\mathbf{x})\right)\right] \le \frac{\theta^2 u^2 U^2}{Bqt^2}.$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i\in\mathcal{U}_p} V_i(\mathbf{x})\right)\left(\sum_{j\in\mathcal{U}_q} V_j(\mathbf{x})\right)\right]\\
=\ & \mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\frac{U}{Bp}\sum_{i\in\mathcal{G}_p} a_t^i\delta_i(\mathbf{x})\right)\left(\frac{U}{Bq}\sum_{j\in\mathcal{G}_q} a_t^j\delta_j(\mathbf{x})\right)\right]\\
\le\ & \frac{\theta^2 u^2 U^2}{Bqt^2}.
\end{aligned}$$

The inequality is by Lemma 8 and Lemma 4. □

**Lemma 7.**

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\frac{1}{j} \le k.$$

*Proof.*

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\frac{1}{j} \le \sum_{i=0}^{k-1}\sum_{j=i+1}^{k}\frac{1}{j} = \sum_i i\times\frac{1}{i} = k.$$

□

**Lemma 8.** *If $p < q$, then*

$$\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{\mathcal{G}_p} a_i\delta_i(\mathbf{x})\right)\left(\sum_{\mathcal{G}_q} b_j\delta_j(\mathbf{x})\right)\right] \le u^2\sum_{\mathcal{G}_p}|a_i||b_i|.$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{\mathcal{G}_p} a_i\delta_i(\mathbf{x})\right)\left(\sum_{\mathcal{G}_q} b_j\delta_j(\mathbf{x})\right)\right]\\
\le\ & \mathbb{E}\left(\sum_{\mathcal{G}_p} a_i b_i\delta_i^2(\mathbf{x}) + \sum_{i\ne j} a_i b_j\delta_i(\mathbf{x})\delta_j(\mathbf{x})\right)\\
\le\ & u^2\sum_{\mathcal{G}_p}|a_i||b_i| + \mathbb{E}\left(\sum_{i\ne j} a_i b_j\delta_i(\mathbf{x})\delta_j(\mathbf{x})\right)
\end{aligned}$$

Note that $\forall i < j$, $\mathbb{E}\left(\delta_i(\mathbf{x})\delta_j(\mathbf{x})\right) = \mathbb{E}_{\mathcal{D}^j,\boldsymbol{\omega}^{j-1}}\left[\mathbb{E}_{\omega^j}\left(\delta_i(\mathbf{x})\delta_j(\mathbf{x})|\boldsymbol{\omega}^{j-1}\right)\right] = \mathbb{E}_{\mathcal{D}^j,\boldsymbol{\omega}^{j-1}}(0) = 0$. Therefore,

$$\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[\left(\sum_{\mathcal{G}_p} a_i\delta_i(\mathbf{x})\right)\left(\sum_{\mathcal{G}_q} b_j\delta_j(\mathbf{x})\right)\right] \le u^2\sum_{\mathcal{G}_p}|a_i||b_i|.$$

□

### A.2 Convergence Rate of CDSG

We analyze the convergence rate of proposed CDSG in Theorem 9. The proof is mainly based on the discussion of Section 4.2.

**Theorem 9 (Convergence rate of CDSG).** *When $\frac{\theta}{t} \le \gamma_t \le \frac{\theta'}{t}$ with $\theta, \theta' > 0$ such that $\theta\lambda \in (1,2) \cup \mathbb{Z}_+$, for any $x \in \mathcal{X}$,*

$$\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left[|f_{t+1}(x) - f_*(x)|^2\right] \leqslant \frac{2C_2^2 + 2\kappa S_2^2}{t},$$

*where*

$$S_2 = \max\left\{\|f_*\|_{\mathcal{H}}, \frac{Q_2 + \sqrt{Q_2^2 + Z(1+\hat{\theta}\lambda)^2\hat{\theta}^2\kappa M^2}}{Z}\right\},$$

*with $Z = 2\lambda\hat{\theta} - 1 Q_2 = \sqrt{2}\kappa^{1/2}LC_2\hat{\theta}$, $C_0 = 2(\kappa+\phi)M\hat{\theta}$, and $\theta \le \hat{\theta} \le \theta'$.*

We first prove by induction to show that

$$\mathbb{E}_{\mathcal{D}^t,\boldsymbol{\omega}^t}\left(|f_{t+1}(\mathbf{x}) - h_{t+1}(\mathbf{x})|^2\right) \le u^2\sum_i^t |a_t^i|^2. \quad (1)$$

Since we are required to sample $\boldsymbol{\omega}_1$ from $\mathbb{P}(\boldsymbol{\omega})$, the base case $t = 1$ holds. For $t > 1$, assume we already sample $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m$, If there is no coordinate $j$ with $\eta_j$ from (9) larger than $\theta/t$, we sample $\boldsymbol{\omega}_t$ from $\mathbb{P}(\boldsymbol{\omega})$. Then the bound holds trivially. The last case is we choose $\boldsymbol{\omega}_k$ as $\boldsymbol{\omega}_t$, which implies 1 holds for the chosen $k$. Then we complete proof for (1). If we set $\eta_k$ as $\theta/t$, by Lemma 4, we get a bound for CDSG exactly the same as the bound for DSG in Theorem 3.

The line search will make the step size larger than $\theta/t$ in the certain iteration $t$. Let the step size of each iteration is $\theta_t/t$, where $\theta_t \leq \theta$. Assume $\theta'$ is the upper bound of all $\theta_t$, then there is $\hat{\theta}$ such that $\sum_{i=1}^{t}(a_t^i)^2 \leqslant \frac{\hat{\theta}^2}{t}$, where $\theta \leq \hat{\theta} \leq \theta'$. In practice, we observe that setting an upper bound on $\theta'$ leads to better performance, since it gives a tighter bound.

## B Experiment with Deep Neural Nets

**CIFAR 10.** We use two convolution layers after contrast normalization and max-pooling layers. We use the feature from the top max-pooling layer from a trained neural net and use PCA to reduce the dimension into 256 for DSG-based algorithms.

**MNIST 8M.** We use LeNet-5 (LeCun et al., 1998) and replace tanh units with rectified linear units. The first two convolutions layers have 16 and 32 filters. The fully connected layer has 128 neurons. We use the features from the last max-pooling layer with dimension 1568 for DSG-based algorithms.

**ImageNet.** We use AlexNet (Krizhevsky et al., 2012) for this dataset. The features for DSG-based algorithms are from the last pooling layer of the jointly-trained neural net.

## References


Dai, B., Xie, B., He, N., Liang, Y., Raj, A., Balcan, M., and Song, L. (2014). Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*.