Supplementary Material for "Gradient Diversity: a Key Ingredient for Scalable Distributed Learning"

1 Examples of Gradient Diversity

1.1 Proof of Remark 1: Generalized linear models

Let $\ell'(\cdot)$ be the derivative of $\ell(\cdot)$. Since we have

$$\nabla f_i(\mathbf{w}) = \ell_i'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\mathbf{x}_i,$$

by letting $a_i := \ell_i'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})$ and $\mathbf{a} = [a_1 \ \cdots \ a_n]^{\mathrm{T}}$, we obtain

$$B_D(\mathbf{w}) = \frac{n\sum_{i=1}^n a_i^2\|\mathbf{x}_i\|_2^2}{\|\sum_{i=1}^n a_i\mathbf{x}_i\|_2^2} = \frac{n\sum_{i=1}^n a_i^2\|\mathbf{x}_i\|_2^2}{\|\mathbf{X}^{\mathrm{T}}\mathbf{a}\|_2^2} \geq \frac{n\min_{i=1,\dots,n}\|\mathbf{x}_i\|_2^2\sum_{i=1}^n a_i^2}{\sigma_{\max}^2(\mathbf{X})\|\mathbf{a}\|_2^2} \geq \frac{n\min_{i=1,\dots,n}\|\mathbf{x}_i\|_2^2}{\sigma_{\max}^2(\mathbf{X})},$$

which completes the proof.

We made a claim after the remark about instantiating it for random design matrices. We provide the proof of that claim below.

1.2 Generalized Linear Function with Random Features

We have the following two results.

Proposition 1. *Suppose that $n \geq d$, and $\mathbf{x}_i$ has i.i.d. σ-sub-Gaussian entries with zero mean. Then, there exist universal constants $c_1, c_2, c_3 > 0$, such that, with probability at least $1 - c_2ne^{-c_3d}$, we have $B_D(\mathbf{w}) \geq c_1d \ \forall \ \mathbf{w} \in \mathcal{W}$.*

Proposition 2. *Suppose that $n \geq d$, and the entries of $\mathbf{x}_i$ are i.i.d. uniformly distributed in $\{-1, 1\}$. Then, there exist universal constants $c_4, c_5, c_6 > 0$, such that, with probability at least $1 - c_5e^{-c_6n}$, we have $B_D(\mathbf{w}) \geq c_4d \ \forall \ \mathbf{w} \in \mathcal{W}$.*

Proof. By the concentration results of the maximum singular value of random matrices, we know that when $n \geq d$, there exist universal constants $C_1, C_2, C_3 > 0$, such that

$$\mathbb{P}\{\sigma_{\max}^2(\mathbf{X}) \leq C_1\sigma^2n\} \geq 1 - C_2e^{-C_3n}. \tag{1}$$

By the concentration results of sub-Gaussian random variables, we know that there exist universal constants $C_4, C_5 > 0$ such that

$$\mathbb{P}\{\|\mathbf{x}_i\|_2^2 \geq C_4\sigma^2d\} \geq 1 - e^{-C_5d},$$

and then by union bound, we have

$$\mathbb{P}\left\{\min_{i=1,\dots,n}\|\mathbf{x}_i\|_2^2 \ge C_4\sigma^2 d\right\} \ge 1 - ne^{-C_5 d}. \tag{2}$$

Then, by combining (1) and (2) and using union bound, we obtain

$$\mathbb{P}\left\{\frac{n\min_{i=1,\dots,n}\|\mathbf{x}_i\|_2^2}{\sigma_{\max}^2(\mathbf{X})} \ge \frac{C_4}{C_1}d\right\} \ge 1 - C_2e^{-C_3 n} - ne^{-C_5 d},$$

which yields the desired result.

Proposition 2 can be proved using the fact that for Rademacher entries, we have $\|\mathbf{x}_i\|_2^2 = d$ with probability one. □

1.3 Proof of Remark 2: Sparse Conflict

We prove the following result for Example 2 in Section 4.1.

Proposition 3. *Let ρ be the maximum degree of all the vertices in G. Then, we have $\forall\ \mathbf{w}\in\mathcal{W}$, $B_D(\mathbf{w}) \ge n/(\rho+1)$.*

Proof. We adopt the convention that when $(i,j)\in E$, we also have $(j,i)\in E$. By definition, we have

$$\begin{aligned}
B_D(\mathbf{w}) &= \frac{n\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2 + \sum_{i\neq j}\langle\nabla f_i(\mathbf{w}),\nabla f_j(\mathbf{w})\rangle} \\
&= \frac{n\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2 + \sum_{(i,j)\in E}\langle\nabla f_i(\mathbf{w}),\nabla f_j(\mathbf{w})\rangle} \\
&\ge \frac{n\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2 + \sum_{(i,j)\in E}\frac{1}{2}\|\nabla f_i(\mathbf{w})\|_2^2 + \frac{1}{2}\|\nabla f_j(\mathbf{w})\|_2^2}.
\end{aligned}$$

Since ρ is the maximum degree of the vertexes in G, we know that for each $i\in[n]$, the term $\frac{1}{2}\|\nabla f_i(\mathbf{w})\|_2^2$ appears at most 2ρ times in the summation $\sum_{(i,j)\in E}\frac{1}{2}\|\nabla f_i(\mathbf{w})\|_2^2 + \frac{1}{2}\|\nabla f_j(\mathbf{w})\|_2^2$. Therefore, we obtain

$$\sum_{(i,j)\in E}\frac{1}{2}\|\nabla f_i(\mathbf{w})\|_2^2 + \frac{1}{2}\|\nabla f_j(\mathbf{w})\|_2^2 \le \rho\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2,$$

which completes the proof. □

2 Convergence Rates

In this section, we prove our convergence results for different types of functions. To assist the demonstration of the proofs of convergence rates, for any $\mathbf{w}\in\mathcal{W}$, we define the following two quantities:

$$M^2(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n\|\nabla f_i(\mathbf{w})\|_2^2 \quad\text{and}\quad G(\mathbf{w}) := \|\nabla F(\mathbf{w})\|_2^2 = \|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\mathbf{w})\|_2^2$$

One can check that the batch-size bound obeys $B_D(\mathbf{w}) = \frac{M^2(\mathbf{w})}{G(\mathbf{w})}$.

2.1 Proof of Lemma 1

We have

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_{(k+1)B}-\mathbf{w}^*\|_2^2 \mid \mathbf{w}_{kB}] =&\mathbb{E}\left[\|\mathbf{w}_{kB}-\mathbf{w}^*-\gamma\sum_{\ell=kB}^{(k+1)B-1}\nabla f_{s_\ell}(\mathbf{w}_{kB})\|_2^2 \mid \mathbf{w}_{kB}\right]\\ =&\|\mathbf{w}_{kB}-\mathbf{w}^*\|_2^2-2\gamma\sum_{\ell=kB}^{(k+1)B-1}\mathbb{E}[\langle\mathbf{w}_{kB}-\mathbf{w}^*,\nabla f_{s_\ell}(\mathbf{w}_{kB})\rangle \mid \mathbf{w}_{kB}]\\ &+\gamma^2\mathbb{E}\left[\|\sum_{\ell=kB}^{(k+1)B-1}\nabla f_{s_\ell}(\mathbf{w}_{kB})\|_2^2 \mid \mathbf{w}_{kB}\right].\end{aligned}$$

Since s_ℓ's are sampled i.i.d. uniformly from $[n]$, we know that

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_{(k+1)B}-\mathbf{w}^*\|_2^2 \mid \mathbf{w}_{kB}] =&\|\mathbf{w}_{kB}-\mathbf{w}^*\|_2^2-2\gamma B\langle\mathbf{w}_{kB}-\mathbf{w}^*,\nabla F(\mathbf{w}_{kB})\rangle\\ &+\gamma^2(BM^2(\mathbf{w}_{kB})+B(B-1)G(\mathbf{w}_{kB}))\\ =&\|\mathbf{w}_{kB}-\mathbf{w}^*\|_2^2-2\gamma B\langle\mathbf{w}_{kB}-\mathbf{w}^*,\nabla F(\mathbf{w}_{kB})\rangle\\ &+\gamma^2 B\left(1+\frac{B-1}{B_D(\mathbf{w}_{kB})}\right)M^2(\mathbf{w}_{kB})\\ =&\|\mathbf{w}_{kB}-\mathbf{w}^*\|_2^2-2\gamma B\langle\mathbf{w}_{kB}-\mathbf{w}^*,\nabla F(\mathbf{w}_{kB})\rangle+\gamma^2 B(1+\delta)M^2(\mathbf{w}_{kB}).\end{aligned} \tag{3}$$

We also mention here that this result becomes inequality for the projected mini-batch SGD algorithm, since Euclidean projection onto a convex set is non-expansive. □

2.2 Proof of Theorem 2

Recall that we have the iteration $\mathbf{w}_{(k+1)B}=\mathbf{w}_{kB}-\gamma\sum_{t=kB}^{(k+1)B-1}\nabla f_{s_t}(\mathbf{w}_{kB})$. Since $F(\mathbf{w})$ has β-Lipschitz gradients, we have

$$F(\mathbf{w}_{(k+1)B})\le F(\mathbf{w}_{kB})+\langle\nabla F(\mathbf{w}_{kB}),\mathbf{w}_{(k+1)B}-\mathbf{w}_{kB}\rangle+\frac{\beta}{2}\|\mathbf{w}_{(k+1)B}-\mathbf{w}_{kB}\|_2^2.$$

Then, we obtain

$$\left\langle\nabla F(\mathbf{w}_{kB}),\gamma\sum_{t=kB}^{(k+1)B-1}\nabla f_{s_t}(\mathbf{w}_{kB})\right\rangle\le F(\mathbf{w}_{kB})-F(\mathbf{w}_{(k+1)B})+\frac{\beta}{2}\left\|\gamma\sum_{t=kB}^{(k+1)B-1}\nabla f_{s_t}(\mathbf{w}_{kB})\right\|_2^2.$$

Now we take expectation on both sides. By iterative expectation, we know that for any $t\ge kB$,

$$\mathbb{E}[\langle\nabla F(\mathbf{w}_{kB}),\nabla f_{s_t}(\mathbf{w}_{kB})\rangle]=\mathbb{E}[\|\nabla F(\mathbf{w}_{kB})\|_2^2].$$

We also have

$$\mathbb{E}\left[\left\|\sum_{t=kB}^{(k+1)B-1}\nabla f_{s_t}(\mathbf{w}_{kB})\right\|_2^2\right]=\mathbb{E}[BM^2(\mathbf{w}_{kB})+B(B-1)G(\mathbf{w}_{kB})]\le B(1+\delta)M^2.$$

Consequently,

$$\gamma B \mathbb{E}[\|\nabla F(\mathbf{w}_{kB})\|_2^2] \leq \mathbb{E}[F(\mathbf{w}_{kB})] - \mathbb{E}[F(\mathbf{w}_{(k+1)B})] + \frac{\beta}{2}\gamma^2 B(1+\delta)M^2. \tag{4}$$

Summing up equation (4) for $k = 0, \ldots, T/B - 1$ yields

$$\gamma B \sum_{k=0}^{T/B-1} \mathbb{E}[\|\nabla F(\mathbf{w}_{kB})\|_2^2] \leq F(\mathbf{w}_0) - F^* + \frac{\beta}{2}\gamma^2 T(1+\delta)M^2,$$

which simplifies to

$$\min_{k=0,\ldots,T/B-1} \mathbb{E}[\|\nabla F(\mathbf{w}_{kB})\|_2^2] \leq \frac{F(\mathbf{w}_0) - F^*}{\gamma T} + \frac{\beta}{2}\gamma(1+\delta)M^2.$$

We can then derive the results by replacing γ and T with the particular choices. □

2.3 Proof of Theorem 3

Substituting $\mathbf{w} = \mathbf{w}_{(k+1)B}$ and $\mathbf{w}' = \mathbf{w}_{kB}$ in the condition for β-smoothness in Definition 1, we obtain

$$F(\mathbf{w}_{(k+1)B}) \leq F(\mathbf{w}_{kB}) - \gamma \left\langle \nabla F(\mathbf{w}_{kB}), \sum_{t=kB}^{(k+1)B-1} \nabla f_{s_t}(\mathbf{w}_{kB}) \right\rangle + \frac{\beta\gamma^2}{2} \left\| \sum_{t=kB}^{(k+1)B-1} \nabla f_{s_t}(\mathbf{w}_{kB}) \right\|_2^2. \tag{5}$$

Condition on $\mathbf{w}_{kB}$ and take expectations over the choice of s_t, $t = kB, \ldots, (k+1)B - 1$. We obtain

$$\mathbb{E}[F(\mathbf{w}_{(k+1)B}) \mid \mathbf{w}_{kB}] \leq F(\mathbf{w}_{kB}) - \gamma B \|\nabla F(\mathbf{w}_{kB})\|_2^2 + \frac{\beta\gamma^2}{2}\left(BM^2(\mathbf{w}_{kB}) + B(B-1)G(\mathbf{w}_{kB})\right). \tag{6}$$

Then, we take expectation over all the randomness of the algorithm. Using the PL condition in Definition 3 and the fact that $B \leq 1 + \delta B_D(\mathbf{w})$ for all $\mathbf{w} \in \mathcal{W}_T$, we write

$$\mathbb{E}\left[F(\mathbf{w}_{(k+1)B}) - F^*\right] \leq (1 - 2\gamma\mu B)\mathbb{E}\left[F(\mathbf{w}_{kB}) - F^*\right] + (1+\delta)\frac{\beta B \gamma^2 M^2}{2}. \tag{7}$$

Then, if $B \leq \frac{1}{2\gamma\mu}$, we have

$$\mathbb{E}\left[F(\mathbf{w}_T) - F^*\right] \leq (1 - 2\gamma\mu B)^{T/B}(F(\mathbf{w}_0) - F^*) + (1+\delta)\frac{\beta\gamma M^2}{4\mu}.$$

Using the fact that $1 - x \leq e^{-x}$ for any $x \geq 0$, and choosing $\gamma = \frac{2\epsilon\mu}{M^2\beta}$, we obtain the desired result. □

2.4 Proof of Theorem 4

According to Lemma 1, for every $k = 0, 1, \dots, \frac{T}{B} - 1$, we have

$$\mathbb{E}[\|\mathbf{w}_{(k+1)B} - \mathbf{w}^*\|_2^2 \mid \mathbf{w}_{kB}] \leq \|\mathbf{w}_{kB} - \mathbf{w}^*\|_2^2 - 2\gamma B \left\langle \nabla F(\mathbf{w}_{kB}), \mathbf{w}_{kB} - \mathbf{w}^* \right\rangle + (1+\delta)\gamma^2 B M^2.$$

Then, we take expectation over all the randomness of the algorithm. Let $D_{kB} = \mathbb{E}[\|\mathbf{w}_{kB} - \mathbf{w}^*\|_2^2]$. We have

$$\mathbb{E}[\langle \nabla F(\mathbf{w}_{kB}), \mathbf{w}_{kB} - \mathbf{w}^* \rangle] \leq \frac{1}{2\gamma B}(D_{kB} - D_{(k+1)B}) + (1+\delta)\frac{\gamma}{2}M^2. \tag{8}$$

We use equation (8) to prove the convergence rate. We have by convexity

$$\begin{aligned}
\mathbb{E}\left[F\left(\frac{B}{T}\sum_{k=0}^{\frac{T}{B}-1}\mathbf{w}_{kB}\right) - F(\mathbf{w}^*)\right] &\leq \mathbb{E}\left[\frac{B}{T}\sum_{k=0}^{\frac{T}{B}-1}F(\mathbf{w}_{kB}) - F(\mathbf{w}^*)\right] \\
&= \frac{B}{T}\sum_{t=0}^{\frac{T}{B}-1}\mathbb{E}[F(\mathbf{w}_{kB}) - F(\mathbf{w}^*)] \\
&\leq \frac{B}{T}\sum_{t=0}^{\frac{T}{B}-1}\mathbb{E}[\langle \nabla F(\mathbf{w}_{kB}), \mathbf{w}_{kB} - \mathbf{w}^* \rangle] \\
&\leq \frac{D_0}{2\gamma T} + (1+\delta)\frac{\gamma M^2}{2},
\end{aligned}$$

where the last inequality is obtained by summing inequality (8) over $k = 0, 1, \dots, \frac{T}{B} - 1$. Then, we can derive the results by replacing γ and T with the particular choices. □

2.5 Proof of Theorem 5

According to Lemma 1, we have

$$\mathbb{E}[\|\mathbf{w}_{(k+1)B} - \mathbf{w}^*\|_2^2 \mid \mathbf{w}_{kB}] \leq \|\mathbf{w}_{kB} - \mathbf{w}^*\|_2^2 - 2\gamma B \left\langle \nabla F(\mathbf{w}_{kB}), \mathbf{w}_{kB} - \mathbf{w}^* \right\rangle + (1+\delta)\gamma^2 B M^2(\mathbf{w}_{kB}).$$

By strong convexity of $F(\mathbf{w})$, we have

$$\langle \nabla F(\mathbf{w}_{kB}), \mathbf{w}_{kB} - \mathbf{w}^* \rangle \geq \lambda \|\mathbf{w}_{kB} - \mathbf{w}^*\|_2^2,$$

which yields

$$\mathbb{E}[\|\mathbf{w}_{(k+1)B} - \mathbf{w}^*\|_2^2 \mid \mathbf{w}_{kB}] \leq (1 - 2\gamma\lambda B)\|\mathbf{w}_{kB} - \mathbf{w}^*\|_2^2 + (1+\delta)\gamma^2 B M^2(\mathbf{w}_{kB}). \tag{9}$$

Then, by taking expectations over the randomness of the whole algorithm on both sizes of (9), we obtain

$$\mathbb{E}[\|\mathbf{w}_{(k+1)B} - \mathbf{w}^*\|_2^2] \leq (1 - 2\gamma\lambda B)\mathbb{E}[\|\mathbf{w}_{kB} - \mathbf{w}^*\|_2^2] + (1+\delta)\gamma^2 B M^2.$$

Then if $B \leq \frac{1}{2\gamma\lambda}$, we obtain

$$\mathbb{E}[\|\mathbf{w}_T - \mathbf{w}^*\|_2^2] \leq (1 - 2\gamma\lambda B)^{T/B}\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 + (1+\delta)\frac{\gamma M^2}{2\lambda}.$$

Using the fact that $1 - x \leq e^{-x}$ for any $x \geq 0$, we obtain

$$\mathbb{E}[\|\mathbf{w}_T - \mathbf{w}^*\|_2^2] \leq e^{-2\gamma\lambda T}D_0 + (1+\delta)\frac{\gamma M^2}{2\lambda}.$$

We complete the proof by taking $\gamma = \frac{\epsilon\lambda}{M^2}$. □

3 Lower Bound

In this section, we prove the lower bound on convergence for strongly convex functions.

3.1 Proof of Theorem 6

We set $f_i(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}-\mathbf{x}_i\|_2^2$, and thus $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \frac{\lambda}{2}\|\mathbf{w}-\mathbf{x}_i\|_2^2$. We choose $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_2 \le 1\}$, and $\mathbf{x}_i$'s such that $\|\mathbf{x}_i\|_2 = 1$ for all $i = 1, \dots, n$, and $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$.

One can check that $\nabla f_i(\mathbf{w}) = \lambda(\mathbf{w} - \mathbf{x}_i)$, $\nabla F(\mathbf{w}) = \lambda\mathbf{w}$, and

$$M^2(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2 = \frac{1}{n}\sum_{i=1}^n \lambda^2\|\mathbf{w} - \mathbf{x}_i\|_2^2 = \frac{1}{n}\sum_{i=1}^n \lambda^2(\|\mathbf{w}\|_2^2 + \|\mathbf{x}_i\|_2^2).$$

Since $M^2(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \lambda^2(\|\mathbf{w}\|_2^2 + \|\mathbf{x}_i\|_2^2) \in [\lambda^2, 2\lambda^2]$ for all $\mathbf{w} \in \mathcal{W}$, we know that we have $M^2(\mathbf{w}) \ge \frac{1}{2}M^2$ for all $\mathbf{w} \in \mathcal{W}$.

Since $\mathcal{W}$ is a bounded set, the projection step has to be taken in order to guarantee that $\mathbf{w}_{N_k} \in \mathcal{W}$. However, one can show that, if the initial guess $\mathbf{w}_0$ is in the convex hull of $\mathbf{x}_1, \dots, \mathbf{x}_n$ (denoted by $\mathcal{C} \subset \mathcal{W}$), then, without using projection, the obtained model parameter $\mathbf{w}_{N_k}$ always stays inside $\mathcal{C}$. More specifically, we have the following result.

Proposition 4. *Suppose that $B_k \le \frac{1}{\lambda\gamma}$ for all $k = 1, \dots, K$, and $\mathbf{w}_0 \in \mathcal{C}$. Then, without using projection, $\mathbf{w}_{N_k} \in \mathcal{C}$ for all k.*

Proof. We prove this result using induction. Suppose that $\mathbf{w}_{N_{k-1}} \in \mathcal{C}$. Then, we have

$$\begin{aligned}\mathbf{w}_{N_k} =& \mathbf{w}_{N_{k-1}} - \gamma\sum_{\ell=N_{k-1}}^{N_k-1} \nabla f_{s_\ell}(\mathbf{w}_{N_{k-1}}) = \mathbf{w}_{N_{k-1}} - \gamma\sum_{\ell=N_{k-1}}^{N_k-1} \lambda(\mathbf{w}_{N_{k-1}} - \mathbf{x}_{s_\ell}) \\ =& (1-\gamma\lambda B_k)\mathbf{w}_{N_{k-1}} + \gamma\lambda B_k\left(\frac{1}{B_k}\sum_{\ell=N_{k-1}}^{N_k-1}\mathbf{x}_{s_\ell}\right).\end{aligned}$$

Since $\mathbf{w}_{N_{k-1}}, \frac{1}{B_k}\sum_{\ell=N_{k-1}}^{N_k-1} \mathbf{x}_{s_\ell} \in \mathcal{C}$, we prove Lemma 4. □

From now on we assume $\mathbf{w}_0 \in \mathcal{C}$ and do not consider projection. According to equation (3) in the proof of Lemma 1, we have[1]

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_{N_k} - \mathbf{w}^*\|_2^2 \mid \mathbf{w}_{N_{k-1}}] =& \|\mathbf{w}_{N_{k-1}} - \mathbf{w}^*\|_2^2 - 2\gamma B_k\langle \mathbf{w}_{N_{k-1}} - \mathbf{w}^*, \nabla F(\mathbf{w}_{N_{k-1}})\rangle \\ &+ \gamma^2 B_k\left(1 + \frac{B_k - 1}{B_D(\mathbf{w}_{N_{k-1}})}\right)M^2(\mathbf{w}_{N_{k-1}}) \\ \ge& (1 - 2\gamma\lambda B_k)\|\mathbf{w}_{N_{k-1}} - \mathbf{w}^*\|_2^2 + \frac{1}{2}\gamma^2 M^2 B_k\left(1 + \frac{B_k - 1}{B_D(\mathbf{w}_{N_{k-1}})}\right).\end{aligned}$$

[1] We still keep $\mathbf{w}^*$ although $\mathbf{w}^* = \mathbf{0}$.

Then, we take expectation over the randomness of the whole algorithm and obtain

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{N_k}-\mathbf{w}^*\|_2^2] \geq &(1-2\gamma\lambda B_k)\mathbb{E}[\|\mathbf{w}_{N_{k-1}}-\mathbf{w}^*\|_2^2]+\frac{1}{2}\gamma^2M^2B_k\left(1+(B_k-1)\mathbb{E}\left[\frac{1}{B_D(\mathbf{w}_{N_{k-1}})}\right]\right)\\
\geq &(1-2\gamma\lambda B_k)\mathbb{E}[\|\mathbf{w}_{N_{k-1}}-\mathbf{w}^*\|_2^2]+\frac{1}{2}\gamma^2M^2B_k\left(1+(B_k-1)\frac{1}{\mathbb{E}[B_D(\mathbf{w}_{N_{k-1}})]}\right)\\
\geq &(1-2\gamma\lambda B_k)\mathbb{E}[\|\mathbf{w}_{N_{k-1}}-\mathbf{w}^*\|_2^2]+\frac{1}{2}(1+\delta)\gamma^2M^2B_k,
\end{aligned}$$

where the second inequality is due to Jensen's inequality, and the third inequality is due to the fact that $B_k \geq 1+\delta\mathbb{E}[B_D(\mathbf{w}_{N_{k-1}})]$.

Rolling out the above recursion, and denoting $\alpha_k = 2\gamma\lambda B_k \in [0,1]$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_{N_K}-\mathbf{w}^*\|_2^2\right] &\geq \|\mathbf{w}_0-\mathbf{w}^*\|_2^2\left(\prod_{k=1}^{K}(1-\alpha_k)\right)+\frac{1}{2}(1+\delta)\gamma^2M^2\left[B_K+\sum_{k=1}^{K-1}\prod_{i=k+1}^{K}(1-\alpha_i)B_k\right]\\
&= \|\mathbf{w}_0-\mathbf{w}^*\|_2^2\left(\prod_{i=1}^{K}(1-\alpha_i)\right)+\frac{1}{4}(1+\delta)\frac{\gamma M^2}{\lambda}\left[\alpha_K+\sum_{k=1}^{K-1}\prod_{i=k+1}^{K}(1-\alpha_i)\alpha_k\right].
\end{aligned}$$

Now the number of gradient updates is given by $\sum_{k=1}^{K} B_k = T$, and consequently, $\sum_{k=1}^{K}\alpha_k = 2\gamma\lambda T$. Since we consider the case when $T \geq \frac{c}{\gamma\lambda}$ for some universal constant $c>0$ (and SGD only converges in this regime), so we have $\sum_{k=1}^{K}\alpha_k \geq 2c$.

Substituting the value of step-size γ, we see that in order to complete the proof, it suffices to show that the quantity

$$J(\alpha) = \alpha_K + \sum_{k=1}^{K-1}\prod_{i=k+1}^{K}(1-\alpha_i)\alpha_k$$

is lower bounded as $\Omega(1)$. In order to show this, note that $J(\alpha)$ can be equivalently expressed as the CDF of a geometric distribution with non-uniform probabilities of success α_k. We could further see that

$$J(\alpha) = 1-\prod_{k=1}^{K}(1-\alpha_k) \geq 1-\left[\frac{1}{K}\sum_{k=1}^{K}(1-\alpha_k)\right]^K \geq 1-(1-2c/K)^K,$$

and the last term is lower bounded by a constant for all $K \geq 1$. □

A second bound that was implicit in our convergence rates is also sharp, as shown in the following section.

3.2 Necessity of $B \leq \mathcal{O}(\frac{1}{\lambda\gamma})$

In this section, we show that, up to a constant factor, the condition $B \leq \frac{1}{2\gamma\lambda}$ in Theorem 5 and 6, is actually necessary for mini-batch SGD to converge when $F(\mathbf{w})$ is strongly convex. More precisely, we can show that, when $B > \frac{2}{\gamma\lambda}$, mini-batch SGD diverges.

Proposition 5. *Suppose that $F(\mathbf{w})$ is λ-strongly convex. Condition on the model parameter $\mathbf{w}_{kB}$ obtained after k iterations. Suppose that $\mathbf{w}_{kB} - \gamma\sum_{i\in\mathcal{I}}\nabla f_i(\mathbf{w}_{kB}) \in \mathcal{W}$ for all $\mathcal{I} \in [n]^B$. Then, if $B > \frac{2}{\gamma\lambda}$, we have*

$$\mathbb{E}\left[\|\mathbf{w}_{(k+1)B}-\mathbf{w}^*\|_2 \mid \mathbf{w}_{kB}\right] > \|\mathbf{w}_{kB}-\mathbf{w}^*\|_2.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_{(k+1)B} - \mathbf{w}_{kB}\|_2 \mid \mathbf{w}_{kB}\right] &\geq \left\|\mathbb{E}[\mathbf{w}_{(k+1)B} - \mathbf{w}_{kB} \mid \mathbf{w}_{kB}]\right\|_2 \\
&= \gamma \left\| \sum_{t=kB}^{(k+1)B-1} \mathbb{E}[\nabla f_{s_t}(\mathbf{w}_{kB}) \mid \mathbf{w}_{kB}] \right\|_2 \\
&= \gamma B \|\nabla F(\mathbf{w}_{kB})\|_2 \\
&\geq \gamma B \lambda \|\mathbf{w}_{kB} - \mathbf{w}^*\|_2,
\end{aligned}$$

where the first step follows by Jensen's inequality, and the last by strong convexity.

This allows us to conclude that if $B > \frac{2}{\gamma\lambda}$, $\mathbb{E}\left[\|\mathbf{w}_{(k+1)B} - \mathbf{w}_{kB}\|_2 \mid \mathbf{w}_{kB}\right] > 2\|\mathbf{w}_{kB} - \mathbf{w}^*\|_2$. Then, by triangle inequality,

$$\mathbb{E}\left[\|\mathbf{w}_{(k+1)B} - \mathbf{w}^*\|_2 \mid \mathbf{w}_{kB}\right] \geq \mathbb{E}\left[\|\mathbf{w}_{(k+1)B} - \mathbf{w}_{kB}\|_2 \mid \mathbf{w}_{kB}\right] - \|\mathbf{w}_{kB} - \mathbf{w}^*\|_2 > \|\mathbf{w}_{kB} - \mathbf{w}^*\|_2,$$

and thus mini-batch SGD diverges. □

We now turn to showing that various heuristics for SGD are also diversity-inducing.

4 Proof of Theorem 7

For DropConnect, we have

$$\begin{aligned}
B_D^{\mathsf{drop}}(\mathbf{w}) &= n\frac{\sum_{i=1}^n \mathbb{E}[\|\mathbf{D}_i \nabla f_i(\mathbf{w})\|_2^2]}{\mathbb{E}[\|\sum_{i=1}^n \mathbf{D}_i \nabla f_i(\mathbf{w})\|_2^2]} \\
&= \frac{n\sum_{i=1}^n (1-p)\|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n (1-p)\|\nabla f_i(\mathbf{w})\|_2^2 + (1-p)^2 \sum_{j\neq k}\langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle}.
\end{aligned} \tag{10}$$

Recall that

$$B_D(\mathbf{w}) = \frac{n\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2 + \sum_{j\neq k}\langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle},$$

and we can see that for any $\mathbf{w}$ such that $\sum_{j\neq k}\langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle \geq 0$, we must have $B_D(\mathbf{w}) \leq n$. In this case, we have

$$B_D^{\mathsf{drop}}(\mathbf{w}) \geq \frac{n\sum_{i=1}^n (1-p)\|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n (1-p)\|\nabla f_i(\mathbf{w})\|_2^2 + (1-p)\sum_{j\neq k}\langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle} = B_D(\mathbf{w}).$$

On the other hand, if $\sum_{j\neq k}\langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle < 0$, we must have $B_D(\mathbf{w}) > n$, and one can simply check that we also have $B_D^{\mathsf{drop}}(\mathbf{w}) > n$.

For stochastic gradient Langevin dynamics, we have

$$B_D^{\mathsf{sgld}}(\mathbf{w}) = \frac{n\sum_{i=1}^n \mathbb{E}[\|\nabla f_i(\mathbf{w}) + \xi_i\|_2^2]}{\mathbb{E}[\|\sum_{i=1}^n (\nabla f_i(\mathbf{w}) + \xi_i)\|_2^2]} = \frac{n\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2 + n^2 d\sigma^2}{\|\sum_{i=1}^n \nabla f_i(\mathbf{w})\|_2^2 + nd\sigma^2}. \tag{11}$$

Therefore, as long as $B_D(\mathbf{w}) = \frac{n\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2}{\|\sum_{i=1}^n \nabla f_i(\mathbf{w})\|_2^2} \leq n$, we have $B_D^{\mathsf{sgld}}(\mathbf{w}) \geq B_D(\mathbf{w})$. In addition, if $B_D(\mathbf{w}) > n$, then $B_D^{\mathsf{sgld}}(\mathbf{w}) > n$.

For quantization, one can simply check that for any $i \in [n]$, we have $\mathbb{E}[\|Q(\nabla f_i(\mathbf{w}))\|_2^2] = \|\nabla f_i(\mathbf{w})\|_2 \|\nabla f_i(\mathbf{w})\|_1$, and for any $j \neq k$, we have $\mathbb{E}[\langle Q(\nabla f_j(\mathbf{w})), Q(\nabla f_k(\mathbf{w}))\rangle] = \langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle$. Consequently,

$$\begin{aligned} B_D^{\mathsf{quant}}(\mathbf{w}) &= \frac{n\sum_{i=1}^n \mathbb{E}[\|Q(\nabla f_i(\mathbf{w}))\|_2^2]}{\mathbb{E}[\|\sum_{i=1}^n Q(\nabla f_i(\mathbf{w}))\|_2^2]} \\ &= \frac{n\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2 \|\nabla f_i(\mathbf{w})\|_1}{\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2 \|\nabla f_i(\mathbf{w})\|_1 + \sum_{j\neq k} \langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle}. \end{aligned} \tag{12}$$

We define

$$\Delta_D^{\mathsf{quant}}(\mathbf{w}) := \frac{\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2 \|\nabla f_i(\mathbf{w})\|_1}{\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2 \|\nabla f_i(\mathbf{w})\|_1 + \sum_{j\neq k} \langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle},$$

and

$$\Delta_D(\mathbf{w}) := \frac{\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2}{\sum_{i=1}^n \|\nabla f_i(\mathbf{w})\|_2^2 + \sum_{j\neq k} \langle \nabla f_j(\mathbf{w}), \nabla f_k(\mathbf{w})\rangle},$$

and we have $B_D^{\mathsf{quant}}(\mathbf{w}) = n\Delta_D^{\mathsf{quant}}(\mathbf{w})$ and $B_D = n\Delta_D(\mathbf{w})$. One can now check that due to the fact that $\|\mathbf{v}\|_2\|\mathbf{v}\|_1 \geq \|\mathbf{v}\|_2^2$ for any vector $\mathbf{v}$, when $\Delta_D(\mathbf{w}) \in (0,1)$, we have $\Delta_D^{\mathsf{quant}}(\mathbf{w}) > \Delta_D(\mathbf{w})$, and when $\Delta_D(\mathbf{w}) > 1$, we have $\Delta_D^{\mathsf{quant}}(\mathbf{w}) > 1$. □

5 Stability

Let us begin by defining some useful notation. We let

$$\overline{M}^2(\mathbf{w}, \mathbf{w}') := \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}')\|_2^2 \quad \text{and} \quad \overline{G}(\mathbf{w}, \mathbf{w}') := \|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\|_2^2.$$

One can see that $\overline{B}_D(\mathbf{w}, \mathbf{w}') = \frac{\overline{M}^2(\mathbf{w}, \mathbf{w}')}{\overline{G}(\mathbf{w}, \mathbf{w}')}$. We also define

$$\overline{B}_D = \inf_{\mathbf{w}\neq\mathbf{w}'} \overline{B}_D(\mathbf{w}, \mathbf{w}').$$

Before turning to the proofs, we first provide some background.

5.1 Background on Stability and Generalization

Recall that in supervised learning problems, our goal is to learn a parametric model with small population risk $R(\mathbf{w}) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{w};\mathbf{z})]$. In order to do so, we use empirical risk minimization, and hope to obtain a model that has both small empirical risk and small population risk to avoid overfitting. Formally, let A be a possibly randomized algorithm which maps the training data to the parameter space as $\mathbf{w} = A(\mathcal{S})$. We define the *expected generalization error* of the algorithm as

$$\epsilon_{\text{gen}}(A) := \left|\mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(A(\mathcal{S})) - R(A(\mathcal{S}))]\right|.$$

Bousquet and Ellisseef show that there is a fundamental connection between the generalization error and algorithmic stability. An algorithm is said to be stable if it produces similar models given similar training data. We summarize their result as follows.

Proposition 6. *Let $\mathcal{S} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\mathcal{S}' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_n)$ be two independent random samples from $\mathcal{D}$, and let $\mathcal{S}^{(i)} = (\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}'_i, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_n)$ be the sample that is identical to $\mathcal{S}$ except in the i-th data point where we replace $\mathbf{z}_i$ with $\mathbf{z}'_i$. Then, we have*

$$\mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(A(\mathcal{S})) - R(A(\mathcal{S}))] = \mathbb{E}_{\mathcal{S},\mathcal{S}',A}\left[\frac{1}{n}\sum_{i=1}^{n} f(A(\mathcal{S}^{(i)}); \mathbf{z}'_i) - \frac{1}{n}\sum_{i=1}^{n} f(A(\mathcal{S}); \mathbf{z}'_i)\right].$$

With the notation in Proposition 6, we define the following quantity that characterizes the algorithmic *stability* of the learning algorithm given the data points:

$$\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}') = \mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A(\mathcal{S}^{(i)}); \mathbf{z}'_i) - \frac{1}{n}\sum_{i=1}^{n} f(A(\mathcal{S}); \mathbf{z}'_i)\right], \tag{13}$$

where we condition on the data sets $\mathcal{S}$ and $\mathcal{S}'$ and take expectation over the randomness of the learning algorithm (mini-batch SGD). Recall from Theorem 6 that

$$\epsilon_{\text{gen}}(A) = \left|\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right]\right| \le \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\left|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right|\right]. \tag{14}$$

We bound $\epsilon_{\text{gen}}(A)$ by first showing a bound on $\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')$ that depends on the sample $(\mathcal{S}, \mathcal{S}')$, then using equation (14) to obtain, as a corollary, results for generalization error.

For convex and strongly convex functions, we have the following two results on stability.

Proposition 7 (stability of convex functions)**.** *Fix sample $(\mathcal{S}, \mathcal{S}')$. Suppose that for any $\mathbf{z} \in \mathcal{Z}$, $f(\mathbf{w}; \mathbf{z})$ is convex, L-Lipschitz and β-smooth in $\mathcal{W}$. Provided the step-size and batch-size satisfy*

$$\gamma \le \frac{2}{\beta\left(1 + \frac{1}{n-1}\mathbb{1}_{B>1} + \frac{B-1}{B_D(\mathbf{w}, \mathbf{w}')}\right)}, \tag{15}$$

for all $\mathbf{w} \neq \mathbf{w}'$, we have $|\epsilon_{stab}(\mathcal{S}, \mathcal{S}')| \le 2\gamma L^2 \frac{T}{n}$.

Proposition 8 (stability of strongly convex functions)**.** *Fix the sample $(\mathcal{S}, \mathcal{S}')$. Suppose that for any $\mathbf{z} \in \mathcal{Z}$, $f(\mathbf{w}; \mathbf{z})$ is L-Lipschitz, β-smooth, and λ-strongly convex in $\mathcal{W}$, and that $B \le \frac{1}{2\gamma\lambda}$. Provided the step-size and batch-size satisfy*

$$\gamma \le \frac{2}{(\beta + \lambda)\left(1 + \frac{1}{n-1}\mathbb{1}_{B>1} + \frac{B-1}{B_D(\mathbf{w}, \mathbf{w}')}\right)}, \tag{16}$$

for all $\mathbf{w} \neq \mathbf{w}'$, we have $|\epsilon_{stab}(\mathcal{S}, \mathcal{S}')| \le \frac{4L^2}{\lambda n}$.

We also note that Theorem 9 and Theorem 10 in our main paper can be derived as Corollaries of Proposition 7 and 8, respectively. In following sections, we provide the details.

5.2 Proofs of Proposition 7 and Theorem 9

We first recall the problem setting. Suppose that there are two sample sets $\mathcal{S}$ and $\mathcal{S}^{(I)}$ which differ at one data point located at a random position I, which is uniformly distributed in $[n]$. We run the same (projected) parallel mini-batch SGD on both data sets, and after the k-th parallel iteration, we obtain $\mathbf{w}_{kB}$ and $\widetilde{\mathbf{w}}_{kB}$, respectively. After a total number of T gradient updates, *i.e.*,

T/B parallel iterations, we obtain $\mathbf{w}_T$ and $\widetilde{\mathbf{w}}_T$. Let s_t, $t = 0, 1, \ldots, T-1$ be the sequence of indices of samples used by the algorithm. In our setting, s_t are i.i.d. uniformly distributed in $\{1, 2, \ldots, n\}$. Let $\mathbf{z}_{s_t} \in \mathcal{S}$ and $\widetilde{\mathbf{z}}_{s_t} \in \mathcal{S}^{(I)}$, $t = 0, \ldots, T-1$ be the data point used in the algorithms running on the two data sets, respectively. Then, we know that with probability $1 - \frac{1}{n}$, $\mathbf{z}_{s_t} = \widetilde{\mathbf{z}}_{s_t}$, and with probability $\frac{1}{n}$, $\mathbf{z}_{s_t} \neq \widetilde{\mathbf{z}}_{s_t}$. We simplify the notations of the risk function associated with $\mathbf{z}_{s_t}$ and $\widetilde{\mathbf{z}}_{s_t}$ by $f_{s_t}(\mathbf{w}) := f(\mathbf{w}; \mathbf{z}_{s_t})$, and $\widetilde{f}_{s_t}(\mathbf{w}) := f(\mathbf{w}; \widetilde{\mathbf{z}}_{s_t})$, respectively.

We now prove Proposition 7. Throughout this proof, we only consider the case where $B > 1$ and omit the indicator function $\mathbb{1}_{B>1}$. We condition on the data sets and the event that the choice of γ is "good", as shown in (15). Specifically, we condition on the samples $\mathcal{S}$ and $\mathcal{S}'$, and the event Γ:

$$\Gamma = \left\{ \gamma \leq \frac{2}{\beta(1 + \frac{1}{n-1} + \frac{B-1}{\overline{B}_D})} \right\} = \left\{ \overline{B}_D \geq \frac{B-1}{\frac{2}{\gamma\beta} - 1 - \frac{1}{n-1}} \right\}. \tag{17}$$

Recall the definition of η:

$$\eta = \mathbb{P}\left\{ \inf_{\mathbf{w} \neq \mathbf{w}'} \overline{B}_D(\mathbf{w}, \mathbf{w}') < \frac{B-1}{\frac{2}{\gamma\beta} - 1 - \frac{1}{n-1}\mathbb{1}_{B>1}} \right\}. \tag{18}$$

We know that $\eta = \mathbb{P}\{\bar{\Gamma}\}$, and so our goal is to bound $|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')|$ conditioned on the event Γ. Since we assume that $f(\mathbf{w}; \mathbf{z})$ is L-Lipschitz on $\mathcal{W}$, we have

$$\left|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right| \leq L\mathbb{E}_{I,A|\Gamma}\left[\|A(\mathcal{S}^{(I)}) - A(\mathcal{S})\|_2\right] = L\mathbb{E}_{I,A|\Gamma}\left[\|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2\right], \tag{19}$$

and thus it suffices to bound $\mathbb{E}_{I,A|\Gamma}\left[\|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2\right]$.

Consider the samples used in the $(k+1)$-th parallel iteration in the two algorithm instances, *i.e.*, $\{\mathbf{z}_{s_t}\}_{t=kB}^{(k+1)B-1}$, and $\{\widetilde{\mathbf{z}}_{s_t}\}_{t=kB}^{(k+1)B-1}$. Let H_{k+1} be the the number of samples that differ between the two minibatches in iteration $k+1$. According to our sampling scheme, $H_{k+1} \sim \text{bin}(B, \frac{1}{n})$. We condition on the event that $H_{k+1} = h$. Without loss of generality, we assume that $\mathbf{z}_{s_t} = \widetilde{\mathbf{z}}_{s_t}$ for all $t = kB, \ldots, (k+1)B - h - 1$, and $\mathbf{z}_{s_t} \neq \widetilde{\mathbf{z}}_{s_t}$ for all $t = (k+1)B - h, \ldots, (k+1)B - 1$. Consider the first $B - h$ terms. For the unconstrained optimization, we have

$$\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2 = \|(\mathbf{w}_{kB} - \gamma \sum_{t=kB}^{(k+1)B-h-1} \nabla f_{s_t}(\mathbf{w}_{kB})) - (\widetilde{\mathbf{w}}_{kB} - \gamma \sum_{t=kB}^{(k+1)B-h-1} \nabla \widetilde{f}_{s_t}(\widetilde{\mathbf{w}}_{kB}))\|_2^2. \tag{20}$$

For the algorithm with projection, the B gradient update steps are the same as the unconstrained algorithm, and projection step is conducted once all the gradient updates are finished. Therefore, equation (20) also holds for projected algorithm.

Since $f_{s_t}(\mathbf{w}) = \widetilde{f}_{s_t}(\mathbf{w})$ for all $t = kB, \dots, (k+1)B - h - 1$, we further have

$$\begin{aligned}
&\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2 \\
=&\|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - 2\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \gamma \sum_{t=kB}^{(k+1)B-h-1} \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle \\
&+ \gamma^2 \| \sum_{t=kB}^{(k+1)B-h-1} \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\|_2^2 \\
=&\|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - 2\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \gamma \sum_{t=kB}^{(k+1)B-h-1} \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle \\
&+ \gamma^2 \sum_{t=kB}^{(k+1)B-h-1} \|\nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\|_2^2 \\
&+ 2\gamma^2 \sum_{i=kB}^{(k+1)B-h-1} \sum_{j=i+1}^{(k+1)B-h-1} \langle \nabla f_{s_i}(\mathbf{w}_{kB}) - \nabla f_{s_i}(\widetilde{\mathbf{w}}_{kB}), \nabla f_{s_j}(\mathbf{w}_{kB}) - \nabla f_{s_j}(\widetilde{\mathbf{w}}_{kB})\rangle.
\end{aligned} \tag{21}$$

We denote the sequence of indices selected by the mini-batch SGD algorithm up to the t-th sampled data point by A_t, i.e., $A_t = \{s_0, \dots, s_{t-1}\}$. In the following steps, we condition on A_{kB} and the event that $H_{k+1} = h$, and take expectation over the randomness of the SGD algorithm in the $(k+1)$-th parallel iteration and the random choice of I.

We consider each term in equation (21). For the term $\|\nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\|_2^2$, conditioned on the event that $\mathbf{z}_{s_t} = \widetilde{\mathbf{z}}_{s_t}$, we know that s_t is uniformly distributed in $[n]\backslash\{I\}$. Since I is uniformly distributed in $[n]$, we know that the marginal distribution of s_t is uniform on the set $[n]$. We have

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\|_2^2] = \overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}).$$

Then, we find the conditional expectation of $\langle \nabla f_{s_i}(\mathbf{w}_{kB}) - \nabla f_{s_i}(\widetilde{\mathbf{w}}_{kB}), \nabla f_{s_j}(\mathbf{w}_{kB}) - \nabla f_{s_j}(\widetilde{\mathbf{w}}_{kB})\rangle$. The following lemma does precisely this.

Proposition 9. *For any i, j such that $kB \le i, j \le (k+1)B - h - 1$ and $i \neq j$, we have*

$$\begin{aligned}
&\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\langle \nabla f_{s_i}(\mathbf{w}_{kB}) - \nabla f_{s_i}(\widetilde{\mathbf{w}}_{kB}), \nabla f_{s_j}(\mathbf{w}_{kB}) - \nabla f_{s_j}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad = \frac{1}{(n-1)^2}\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}) + \frac{n(n-2)}{(n-1)^2}\overline{G}(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}).
\end{aligned} \tag{22}$$

We prove Proposition 9 in Appendix 5.4. Combining this lemma with the result of equation (21), we have

$$\begin{aligned}
&\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2] \\
&\quad= \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - 2\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \gamma \sum_{t=kB}^{(k+1)B-h-1} \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad+ \gamma^2(B-h)\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}) \\
&\quad+ \gamma^2(B-h)(B-h-1)\left[\frac{1}{(n-1)^2}\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}) + \frac{n(n-2)}{(n-1)^2}\overline{G}(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB})\right] \\
&\quad\le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - 2\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \gamma \sum_{t=kB}^{(k+1)B-h-1} \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad+ \gamma^2(B-h)\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}) + \gamma^2(B-h)\left[\frac{1}{n-1}\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}) + (B-1)\frac{\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB})}{\overline{B}_D(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB})}\right] \\
&\quad\le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - 2\gamma \sum_{t=kB}^{(k+1)B-h-1} \mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad+ \gamma^2(B-h)(1 + \frac{1}{n-1} + \frac{B-1}{\overline{B}_D})\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}).
\end{aligned} \tag{23}$$

By the co-coercive property of convex and smooth functions, we know that

$$\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle \ge \frac{1}{\beta}\|\nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\|_2^2.$$

We thus obtain

$$\begin{aligned}
&\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2] \\
&\quad\le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - (2\frac{\gamma}{\beta} - \gamma^2(1 + \frac{1}{n-1} + \frac{B-1}{\overline{B}_D}))(B-h)\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}).
\end{aligned} \tag{24}$$

Since we condition on the event Γ, we have that γ obeys the relation in equation (17). Consequently,

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2] \le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2.$$

Then by Jensen's inequality, we have

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2] \le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2. \tag{25}$$

For the last h terms, since the loss functions are all L-Lipschitz, we obtain

$$\|\mathbf{w}_{(k+1)B} - \widetilde{\mathbf{w}}_{(k+1)B}\|_2 \le \|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2 + 2\gamma Lh. \tag{26}$$

Then, combining with equation (25), we have

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B} - \widetilde{\mathbf{w}}_{(k+1)B}\|_2] \le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2 + 2\gamma Lh. \tag{27}$$

Recall that H_{k+1} is a binomial random variable. Taking expectation over H_{k+1} yields

$$\mathbb{E}_{I,A|A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B} - \widetilde{\mathbf{w}}_{(k+1)B}\|_2] \le \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2 + 2\gamma L\frac{B}{n}.$$

Then we take expectation over the randomness of the first k parallel iterations and obtain

$$\mathbb{E}_{I,A|\Gamma}[\|\mathbf{w}_{(k+1)B} - \widetilde{\mathbf{w}}_{(k+1)B}\|_2] \leq \mathbb{E}_{I,A|\Gamma}[\|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2] + 2\gamma L \frac{B}{n}. \tag{28}$$

Summing up (28) for $k = 0, 1, \ldots, \frac{T}{B} - 1$ and taking expectation over the data sets, we have

$$\mathbb{E}_{I,A|\Gamma}[\|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2] \leq 2\gamma L \frac{T}{n}. \tag{29}$$

Combining equations (19) and (29), we complete the proof of Proposition 7, *i.e.,* when the event Γ occurs, we have

$$\left|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right| \leq L\mathbb{E}_{I,A|\Gamma}\left[\|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2\right] \leq 2\gamma L^2 \frac{T}{n}. \tag{30}$$

To prove Theorem 9, we notice that when Γ does not occur, we simply have

$$\left|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right| \leq L\mathbb{E}_{I,A|\bar{\Gamma}}[\|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2] \leq 2\gamma L^2 T. \tag{31}$$

Using equations (30) and (31) along with the definition of η, we obtain

$$\begin{aligned}
\epsilon_{\text{gen}} &\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'|\Gamma}\left[\left|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right|\right]\mathbb{P}\{\Gamma\} + \mathbb{E}_{\mathcal{S},\mathcal{S}'|\bar{\Gamma}}\left[\left|\epsilon_{\text{stab}}(\mathcal{S}, \mathcal{S}')\right|\right]\mathbb{P}\{\bar{\Gamma}\} \\
&\leq 2\gamma L^2 \frac{T}{n}(1 - \eta) + 2\gamma L^2 T\eta,
\end{aligned}$$

which completes the proof. □

5.3 Proof of Proposition 8 and Theorem 10

The proof of Proposition 8 follows an argument similar to the proof of Proposition 7. We define the analogous event Γ to signal that the step size is "good", according to equation (16); note that this is slightly different from convex risk functions:

$$\Gamma = \left\{\gamma \leq \frac{2}{(\beta + \lambda)(1 + \frac{1}{n-1} + \frac{B-1}{\overline{B}_D})}\right\} = \left\{\overline{B}_D \geq \frac{B-1}{\frac{2}{\gamma(\beta+\lambda)} - 1 - \frac{1}{n-1}}\right\}. \tag{32}$$

Recall the definition of η:

$$\eta = \mathbb{P}\left\{\inf_{\mathbf{w} \neq \mathbf{w}'} \overline{B}_D(\mathbf{w}, \mathbf{w}') < \frac{B-1}{\frac{2}{\gamma(\beta+\lambda)} - 1 - \frac{1}{n-1}\mathbb{1}_{B>1}}\right\}, \tag{33}$$

we know that $\eta = \mathbb{P}\{\bar{\Gamma}\}$. To prove Proposition 8, our goal is still to bound $\mathbb{E}_{I,A|\Gamma}\left[\|\mathbf{w}_T - \widetilde{\mathbf{w}}_T\|_2\right]$. Since the result in (23) still holds for strongly convex functions, we have

$$\begin{aligned}
&\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h} - \widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2] \\
&\quad \leq \|\mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}\|_2^2 - 2\gamma \sum_{t=kB}^{(k+1)B-h-1} \mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\langle \mathbf{w}_{kB} - \widetilde{\mathbf{w}}_{kB}, \nabla f_{s_t}(\mathbf{w}_{kB}) - \nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad + \gamma^2(B - h)(1 + \frac{1}{n-1} + \frac{B-1}{\overline{B}_D})\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}),
\end{aligned} \tag{34}$$

where H_{k+1} is defined in the same way as in the proof of Proposition 7. For strongly convex functions, we have the following co-coercive property:

$$\langle \mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}, \nabla f_{s_t}(\mathbf{w}_{kB})-\nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\rangle \geq \frac{\beta\lambda}{\beta+\lambda}\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2^2+\frac{1}{\beta+\lambda}\|\nabla f_{s_t}(\mathbf{w}_{kB})-\nabla f_{s_t}(\widetilde{\mathbf{w}}_{kB})\|_2^2,$$

which gives us

$$\begin{aligned}&\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h}-\widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2]\\ &\quad\leq \left(1-2\gamma(B-h)\frac{\beta\lambda}{\beta+\lambda}\right)\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2^2-\gamma(B-h)\left[\frac{2}{\beta+\lambda}-\gamma(1+\frac{1}{n-1}+\frac{B-1}{\overline{B}_D})\right]\overline{M}^2(\mathbf{w}_{kB},\widetilde{\mathbf{w}}_{kB}).\end{aligned} \tag{35}$$

Since we only consider the regime where $B \leq \frac{1}{2\gamma\lambda}$, one can check that $1-2\gamma(B-h)\frac{\beta\lambda}{\beta+\lambda}>0$ for any $h=0,\ldots,B$. Conditioned on the data sets and the event Γ, we have

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h}-\widetilde{\mathbf{w}}_{(k+1)B-h}\|_2^2]\leq\left(1-2\gamma B\frac{\beta\lambda}{\beta+\lambda}\right)\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2^2. \tag{36}$$

With Jensen's inequality and the fact that $\sqrt{1-x}\leq 1-\frac{x}{2}$ for any $x\in[0,1]$, we have

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B-h}-\widetilde{\mathbf{w}}_{(k+1)B-h}\|_2]\leq\left(1-\gamma B\frac{\beta\lambda}{\beta+\lambda}\right)\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2. \tag{37}$$

For the last h terms, we have

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B}-\widetilde{\mathbf{w}}_{(k+1)B}\|_2]\leq\|\mathbf{w}_{(k+1)B-h}-\widetilde{\mathbf{w}}_{(k+1)B-h}\|_2+2\gamma Lh. \tag{38}$$

Combined with equation (37), we obtain

$$\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B}-\widetilde{\mathbf{w}}_{(k+1)B}\|_2]\leq\left(1-\gamma B\frac{\beta\lambda}{\beta+\lambda}\right)\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2+2\gamma Lh,$$

and by taking expectation over h we have

$$\mathbb{E}_{I,A|A_{kB},\Gamma}[\|\mathbf{w}_{(k+1)B}-\widetilde{\mathbf{w}}_{(k+1)B}\|_2]\leq\left(1-\gamma B\frac{\beta\lambda}{\beta+\lambda}\right)\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2+2\gamma L\frac{B}{n}.$$

Taking expectation over A_{kB} yields

$$\mathbb{E}_{I,A|\Gamma}[\|\mathbf{w}_{(k+1)B}-\widetilde{\mathbf{w}}_{(k+1)B}\|_2]\leq\left(1-\gamma B\frac{\beta\lambda}{\beta+\lambda}\right)\mathbb{E}_{I,A|\Gamma}[\|\mathbf{w}_{kB}-\widetilde{\mathbf{w}}_{kB}\|_2]+2\gamma L\frac{B}{n}. \tag{39}$$

Iterating equation (39) yields

$$\mathbb{E}_{I,A|\Gamma}[\|\mathbf{w}_T-\widetilde{\mathbf{w}}_T\|_2]\leq\frac{4L}{\lambda n}. \tag{40}$$

Combining equations (19) and (40), we prove Proposition 8, *i.e.,* when Γ occurs,

$$\left|\epsilon_{\text{stab}}(\mathcal{S},\mathcal{S}')\right|\leq L\mathbb{E}_{I,A|\Gamma}\left[\|\mathbf{w}_T-\widetilde{\mathbf{w}}_T\|_2\right]\leq\frac{4L^2}{\lambda n}. \tag{41}$$

To prove Theorem 10, we notice the fact that, when Γ does not occur, we simply have

$$\left|\epsilon_{\text{stab}}(\mathcal{S},\mathcal{S}')\right|\leq L\mathbb{E}_{I,A|\bar{\Gamma}}[\|\mathbf{w}_T-\widetilde{\mathbf{w}}_T\|_2]\leq 2\gamma L^2T. \tag{42}$$

Combining equations (41) and (42) with the definition of η then yields

$$\begin{aligned}\epsilon_{\text{gen}} &\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'|\Gamma}\left[\left|\epsilon_{\text{stab}}(\mathcal{S},\mathcal{S}')\right|\right]\mathbb{P}\{\Gamma\} + \mathbb{E}_{\mathcal{S},\mathcal{S}'|\bar{\Gamma}}\left[\left|\epsilon_{\text{stab}}(\mathcal{S},\mathcal{S}')\right|\right]\mathbb{P}\{\bar{\Gamma}\} \\ &\leq \frac{4L^2}{\lambda n}(1-\eta) + 2\gamma L^2 T\eta,\end{aligned}$$

which completes the proof. □

5.4 Proof of Proposition 9

One can interpret $\overline{M}^2(\mathbf{w},\mathbf{w}')$ and $\overline{G}(\mathbf{w},\mathbf{w}')$ as follows. Let $\mathcal{P}_1$ be a distribution on $[n]\times[n]$ with PMF

$$p_1(u,v) = \frac{1}{n}\mathbb{1}_{u=v}, \tag{43}$$

and $\mathcal{P}_2$ be the uniform distribution on $[n]\times[n]$, i.e.,

$$p_2(u,v) = \frac{1}{n^2} \tag{44}$$

for all $(u,v)\in[n]\times[n]$. Then, we know that

$$\overline{M}^2(\mathbf{w},\mathbf{w}') = \mathbb{E}_{(i,j)\sim\mathcal{P}_1}[\langle\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}'), \nabla f_j(\mathbf{w}) - \nabla f_j(\mathbf{w}')\rangle],$$

and

$$\overline{G}(\mathbf{w},\mathbf{w}') = \mathbb{E}_{(i,j)\sim\mathcal{P}_2}[\langle\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}'), \nabla f_j(\mathbf{w}) - \nabla f_j(\mathbf{w}')\rangle].$$

Then we find the joint distribution $\mathcal{P}_3$ of (s_i,s_j) where $kB\leq i,j\leq(k+1)B-h-1$ and $i\neq j$. Since $\mathbf{z}_{s_t} = \widetilde{\mathbf{z}}_{s_t}$, we know that $s_t\neq I$ for all $t = kB,\ldots,(k+1)B-h-1$. Then conditioned on I, (s_i,s_j) is uniformly distributed in $([n]\setminus\{I\})\times([n]\setminus\{I\})$. For any $u\in[n]$, we have

$$\begin{aligned}p_3(u,u) = \mathbb{P}\{s_i = u, s_j = u\} &= \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{P}\{s_i = u, s_j = u \mid I = \ell\} \\ &= \frac{1}{n}\sum_{\ell=u}\mathbb{P}\{s_i = u, s_j = u \mid I = \ell\} = \frac{1}{n(n-1)}.\end{aligned}$$

For any $(u,v)\in[n]\times[n]$ such that $u\neq v$, we have

$$\begin{aligned}p_3(u,v) = \mathbb{P}\{s_i = u, s_j = v\} &= \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{P}\{s_i = u, s_j = v \mid I = \ell\} \\ &= \frac{1}{n}\sum_{\ell\neq u,v}\mathbb{P}\{s_i = u, s_j = v \mid I = \ell\} \\ &= \frac{n-2}{n(n-1)^2}.\end{aligned}$$

Then, we know that

$$p_3(u,v) = \frac{1}{(n-1)^2}p_1(u,v) + \frac{n(n-2)}{(n-1)^2}p_2(u,v).$$

Therefore, for any i, j such that $kB \le i, j \le (k+1)B - h - 1$ and $i \neq j$, we have

$$\begin{aligned}
&\mathbb{E}_{I,A|H_{k+1},A_{kB},\Gamma}[\langle \nabla f_{s_i}(\mathbf{w}_{kB}) - \nabla f_{s_i}(\widetilde{\mathbf{w}}_{kB}), \nabla f_{s_j}(\mathbf{w}_{kB}) - \nabla f_{s_j}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad = \mathbb{E}_{(s_i,s_j)\sim\mathcal{P}_3}[\langle \nabla f_{s_i}(\mathbf{w}_{kB}) - \nabla f_{s_i}(\widetilde{\mathbf{w}}_{kB}), \nabla f_{s_j}(\mathbf{w}_{kB}) - \nabla f_{s_j}(\widetilde{\mathbf{w}}_{kB})\rangle] \\
&\quad = \frac{1}{(n-1)^2}\overline{M}^2(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}) + \frac{n(n-2)}{(n-1)^2}\overline{G}(\mathbf{w}_{kB}, \widetilde{\mathbf{w}}_{kB}).
\end{aligned}$$

□

5.5 Examples of Differential Gradient Diversity and Diversity-inducing Mechanisms

We now prove auxiliary results for differential gradient diversity that were stated in the main paper.

Generalized Linear Functions We can show that for generalized linear functions, the lower bound in Theorem 1 still holds, *i.e.,* , for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $\mathbf{w} \neq \mathbf{w}'$, we have

$$\overline{B}_D(\mathbf{w}, \mathbf{w}') \ge \frac{\min_{i=1,\dots,n} \|\mathbf{x}_i\|_2^2}{\sigma_{\max}^2(\mathbf{X})}.$$

To see this, one can simply replace $\nabla f_i(\mathbf{w})$ with $\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}')$ in Appendix 1.1, and define $a_i = \ell_i'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - \ell_i'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}')$. The same arguments in Appendix 1.1 still go through. Consequently, for i.i.d. σ-sub-Gaussian features, we have $\overline{B}_D(\mathbf{w}, \mathbf{w}') \ge c_1 d \ \forall\ \mathbf{w}, \mathbf{w}' \in \mathcal{W}$ with probability at least $1 - c_2 n e^{-c_3 d}$; and for Rademacher entries, we have $\overline{B}_D(\mathbf{w}, \mathbf{w}') \ge c_4 d \ \forall\ \mathbf{w}, \mathbf{w}' \in \mathcal{W}$ with probability greater than $1 - c_5 e^{-c_6 n}$.

Sparse Conflicts The result for gradient diversity still holds for $\overline{B}_D(\mathbf{w}, \mathbf{w}')$, *i.e.,* for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $\overline{B}_D(\mathbf{w}, \mathbf{w}') \ge n/(\rho + 1)$, where ρ is the maximum degree of all the vertices in the conflict graph G. To see this, one should notice that the support of $\nabla f_i(\mathbf{w})$ only depends on the data point, instead of the model parameter, and thus, in general, $\nabla f_i(\mathbf{w})$ and $\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}')$ have the same support. Then, one can simply replace $\nabla f_i(\mathbf{w})$ with $\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}')$ in Appendix 1.3 and the same arguments still go through.

DropConnect When we analyze the stability of mini-batch SGD, we apply the *same* algorithm to two different samples $\mathcal{S}$ and $\mathcal{S}^{(I)}$ that only differ at one data point. Since the algorithm is the same, the random matrices $\mathbf{D}_1, \dots, \mathbf{D}_n$ are also the same in the two instances. Therefore, one can replace $\nabla f_i(\mathbf{w})$ with $\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}')$, and the same arguments still work. Then, we know that when $\overline{B}_D(\mathbf{w}, \mathbf{w}') \le n$, we have $\overline{B}_D^{\mathsf{drop}}(\mathbf{w}, \mathbf{w}') \ge \overline{B}_D(\mathbf{w}, \mathbf{w}')$, and when $\overline{B}_D(\mathbf{w}, \mathbf{w}') > n$, we have $\overline{B}_D^{\mathsf{drop}}(\mathbf{w}, \mathbf{w}') > n$.

Stochastic Gradient Langevin Dynamics For SGLD, we can make similar arguments as in dropout, since the additive noise vectors $\xi_1, \dots, \xi_n$ are the same for the two instances. One can then show that when $\overline{B}_D(\mathbf{w}, \mathbf{w}') \le n$, we have $\overline{B}_D^{\mathsf{sgld}}(\mathbf{w}, \mathbf{w}') \ge \overline{B}_D(\mathbf{w}, \mathbf{w}')$, and when $\overline{B}_D(\mathbf{w}, \mathbf{w}') > n$, we have $\overline{B}_D^{\mathsf{sgld}}(\mathbf{w}, \mathbf{w}') > n$. □