A. Lower Bounds for One-shot Parameter Averaging

A.1. Proof of Theorem 1

Before providing the proof details, let us first describe the high-level intuition of our construction. Roughly speaking, one-shot averaging works well when the bias of the predictor returned by each machine (as a random vector in Euclidean space, based on the sampled training data) is much smaller than the variance. Since each such predictor is based on independent data, averaging m such predictors reduces the variance by a factor of m, leading to good guarantees. However, averaging has no effect on the bias, so this method is ineffectual when the bias dominates the variance. The construction below shows that when the strong convexity parameter is small, this can indeed happen.

More specifically, when the strong convexity parameter is smaller than $\mathcal{O}(1/\sqrt{n})$, the magnitude of the deviations of the (random) predictor returned by each machine does not decay with the sample size n. Moreover, its distribution is highly dependent on the data distribution and the shape of f, and is biased in general. Below we use one such construction, which we found to be convenient for precise analytic calculations, but the intuition applies much more broadly.

Specifically, let $\mathcal{W} = [-2/\lambda, \log(1/\lambda)]$, and define the loss function $f(w;z)$ as

$$f(w;z) = \lambda\left(\frac{1}{2}w^2 + \exp(w)\right) - zw.$$

Furthermore, suppose that $z \sim \mathcal{N}(0,1)$, i.e. the examples have a standard Gaussian distribution. Note that this function is λ-strongly convex, and can be shown to satisfy $\mathbb{E}_z[f'(w;z)^2] \leq 9$ for any $w \in \mathcal{W}$.

Let $\hat{w}_1$ be the parameter vector returned by the machine 1 (this is without loss of generality, since all the machines receive examples drawn from the same distribution). The key to the proof is to show that $\hat{w}_1$ is strongly biased, namely that $\mathbb{E}[\hat{w}_1]$ is bounded away from the true optimum w^*. To compute $\mathbb{E}[\hat{w}_1]$, note that $\hat{w}_1$ minimizes the random function

$$\frac{1}{n}\sum_{i=1}^{n} f(w;z_i) = \lambda\left(\frac{1}{2}w^2 + \exp(w)\right) - \frac{\tilde{z}}{\sqrt{n}}w, \tag{21}$$

where $\tilde{z} = (z_1 + \ldots + z_n)/\sqrt{n}$. Note that since $z_1,\ldots,z_n$ are i.i.d. Gaussians, $\tilde{z}$ also has the same Gaussian distribution $\mathcal{N}(0,1)$.

Taking the derivative, equating to zero and slightly manipulating the result, we get that

$$\lambda\sqrt{n}\left(w + \exp(w)\right) = \tilde{z}.$$

The function on the left-hand-side is strictly monotonically increasing, and has a range $[-\infty,\infty]$. Thus, for any $\tilde{z}$, there exists a unique root $w(\tilde{z})$. Moreover, as long as $|\tilde{z}| \leq \sqrt{n}$, it's easy to verify that $w(\tilde{z})$ is within our domain $\mathcal{W} = [-2/\lambda, \log(1/\lambda)]$, hence $\hat{w}_1 = w(\tilde{z})$. Therefore, letting $p(\cdot)$ denote the standard gaussian distribution of $\tilde{z}$, we have

$$\begin{aligned}\mathbb{E}[\hat{w}_1] &= \int_{x=-\sqrt{n}}^{\sqrt{n}} \hat{w}_1 p(x)dx + \int_{|x|>\sqrt{n}} \hat{w}_1 p(x)dx \leq \int_{x=-\sqrt{n}}^{\sqrt{n}} w(x)p(x)dx + \frac{2}{\lambda}\Pr(|\tilde{z}| \geq \sqrt{n}) \\ &\leq \int_{x=0}^{\sqrt{n}} \left(w(x) + w(-x)\right)p(x)dx + \frac{2}{\lambda}\sqrt{\frac{1}{n\exp(n)}},\end{aligned} \tag{22}$$

where in the last step we used the symmetry of the distribution of $\tilde{z}$ and a standard Gaussian tail bound.

We now turn to analyze $w(x) + w(-x)$. First, we have by definition

$$w(x) + \exp(w(x)) = \frac{x}{\lambda\sqrt{n}} \tag{23}$$

for all x, and therefore

$$w(x) + w(-x) = \left(\frac{x}{\lambda\sqrt{n}} - \exp(w(x))\right) + \left(-\frac{x}{\lambda\sqrt{n}} - \exp(w(-x))\right) \leq -\exp(w(x)). \tag{24}$$

Therefore, we have $w(x)+w(-x)\leq 0$ for all x. More precisely, considering (23) and the fact that its left hand size is monotonic in $w(x)$, it's easy to verify that for any $x\geq 0$, we have $w(x)\geq \log\left(\frac{x}{\lambda\sqrt{n}}-\log\left(\frac{x}{\lambda\sqrt{n}}\right)\right)$, and $w(-x)\leq -\frac{x}{\lambda\sqrt{n}}$, so using (24),

$$w(x)+w(-x)\leq -\exp(w(x))\leq -\frac{x}{\lambda\sqrt{n}}+\log\left(\frac{x}{\lambda\sqrt{n}}\right).$$

Since $\log(a)<a/2$ for all $a\geq 0$, this expression is at most $-\frac{x}{2\lambda\sqrt{n}}$. Plugging this back into (22), and using the assumption $n\geq 9$, we get that

$$\mathbb{E}[\hat{w}_1]\leq -\frac{1}{2\lambda\sqrt{n}}\int_{x=0}^{\sqrt{n}} xp(x)dx+\frac{2}{\lambda}\sqrt{\frac{1}{n\exp(n)}}\leq -\frac{1}{2\lambda\sqrt{n}}\int_{x=0}^{\sqrt{9}} xp(x)dx+\frac{2}{\lambda}\sqrt{\frac{1}{n\exp(n)}}.$$

Since $p(x)$ is the standard Gaussian distribution, it can be numerically checked that this is at most

$$-\frac{0.19}{\lambda\sqrt{n}}+\frac{2}{\lambda}\sqrt{\frac{1}{n\exp(n)}}=\frac{1}{\lambda\sqrt{n}}\left(-0.19+2\sqrt{\exp(-n)}\right)\leq -\frac{1}{6\lambda\sqrt{n}}.$$

So, we finally get $\mathbb{E}[\hat{w}_1]\leq -1/(6\lambda\sqrt{n})$.

Now, we show that this expected value of $\hat{w}_1$ is far away from w^*. w^* is not hard to calculate: It satisfies

$$w^*+\exp(w^*)=0,$$

and it can be calculated numerically that $w^*=-0.5671..>-3/5$. Moreover, we assume $\lambda\leq 1/(9\sqrt{n})$, so $\lambda\sqrt{n}\leq 1/9$ and thus it can be verified that

$$w^*-\mathbb{E}[\hat{w}_1]>-\frac{3}{5}+\frac{1}{6\lambda\sqrt{n}}\geq\frac{1}{10\,\lambda\sqrt{n}}.$$

Note that this is always a positive quantity. As a result, using Jensen's inequality, we get

$$\mathbb{E}[(w^*-\bar{w})^2]\geq(w^*-\mathbb{E}[\bar{w}])^2=(w^*-\mathbb{E}[\hat{w}_1])^2\geq\frac{1}{100\,\lambda^2 n}.$$

Moreover, by λ-strong convexity of f, we have that

$$\mathbb{E}[F(\bar{w})-F(w^*)]\geq\mathbb{E}\left[\frac{\lambda}{2}(\bar{w}-w^*)^2\right]\geq\frac{1}{200\,\lambda n}.$$

Finally, it is known that by performing empirical risk minimization over all $N=nm$ instances, and using the fact that $\mathbb{E}[\|\nabla_w f(w,z)\|^2]$ is bounded by a constant, we get

$$\mathbb{E}[(\hat{w}-w^*)^2]\leq\mathcal{O}\left(\frac{1}{\lambda^2 nm}\right)$$

(see (Zhang et al., 2013)) and

$$\mathbb{E}[F(\hat{w})-F(w^*)]\leq\mathcal{O}\left(\frac{1}{\lambda nm}\right)$$

(see Equation (10)). Combining the four inequalities above gives us the theorem statement.

A.2. Bias Correction Also Fails

In (Zhang et al., 2013), which analyzes one-shot parameter averaging, the authors noticed that the analysis fails for small values of λ, and proposed a modification of the simple averaging scheme, designed to reduce bias issues. Specifically, given a parameter $r\in[0,1]$, each machine subsamples rn examples without replacement from its dataset, and computes the optimum $\hat{w}_{2,k}$ with respect to this subsample. Then, it computes the optimum $\hat{w}_{k,1}$ over the entire dataset, and returns the weighted combination $\hat{w}_k=(\hat{w}_{k,1}-r\hat{w}_{k,2})/(1-r)$. Unfortunately, the analysis still results in lower-order terms with bad dependence on λ, and it's not difficult to extend our construction from Theorem 1 to show that this bias-corrected version of the algorithm still fails (at least, if r is chosen in a fixed manner).

For simplicity, we will only sketch the derivation for a fixed choice of λ given n, namely $\lambda = 1/10\sqrt{n}$, and for $r = 1/2$. Also, we assume for simplicity that $\mathcal{W} = \mathbb{R}$ (to avoid tedious dealings with small Gaussian tails). With this choice, the returned solution becomes $\hat{w}_k = 2\hat{w}_{k,1} - \hat{w}_{k,2}$. The distribution of $\hat{w}_{k,1}$, using the same derivation as in the proof of the theorem, is determined by

$$\hat{w}_{k,1} + \exp(\hat{w}_{k,1}) = \frac{1}{\lambda\sqrt{n}}\tilde{z} = 10\tilde{z}$$

where $\tilde{z}$ has a standard Gaussian distribution. As to $\hat{w}_{k,2}$, its distribution is similar to that of $\hat{w}_{k,1}$ with the same choice of λ but only half as many points, hence

$$\hat{w}_{k,2} + \exp(\hat{w}_{k,2}) = \frac{1}{\lambda\sqrt{n/2}}\tilde{z} = 10\sqrt{2}\,\tilde{z}.$$

By a numerical calculation, one can verify that $\mathbb{E}[\hat{w}_k] = 2\,\mathbb{E}[\hat{w}_{k,1}] - \mathbb{E}[\hat{w}_{k,2}] \approx 2 * (-3.3) - (-4.8) = -1.8$. In contrast, $w^* = -0.5671...$ as discussed in the proof. Thus, the bias is constant and does not scale down with the data size, getting a similar effect as in Theorem 1

B. Proof of Theorem 2

For any $w^{(t-1)}$, the optimal solution is always given by:

$$\hat{w} = \arg\min_{w} \phi(w) = w^{(t-1)} - H^{-1}\nabla\phi(w^{(t-1)}). \tag{25}$$

Following (16), we have:

$$\begin{aligned} w^{(t)} &= w^{(t-1)} - \eta\left(\frac{1}{m}\sum(H_i + \mu I)^{-1}\right)\nabla\phi(w^{(t-1)}) \\ &= w^{(t-1)} - \eta\tilde{H}^{-1}\nabla\phi(w^{(t-1)}). \end{aligned}$$

Therefore

$$w^{(t)} - \hat{w} = (H^{-1} - \eta\tilde{H}^{-1})\nabla\phi(w^{(t-1)}) = (I - \eta\tilde{H}^{-1}H)(w^{(t-1)} - \hat{w}). \tag{26}$$

where for the last equality we rearranged (25) to calculate $\nabla\phi(w^{(t-1)}) = H(w^{(t-1)} - \hat{w})$. Bounding $\|Av\| \leq \|A\|_2\|v\|$ and iterating (26) leads to the desired result.

C. Proof of Lemma 1

We will need two auxiliary lemmas:

Lemma 3. *For any positive definite matrix H:*

$$\|I - (H + \mu I)^{-1}H\| = \frac{\mu}{\lambda + \mu},$$

where λ is the smallest eigenvalue of H.

Proof. Write $H = USU^\top$, then

$$\begin{aligned} &\|I - (H + \mu I)^{-1}H\| = \|UIU^\top - (U(S + \mu I)U^\top)^{-1}USU^\top\| = \|UIU^\top - U(S + \mu I)^{-1}U^\top USU^\top\| \\ &= \|U\left(I - (S + \mu I)^{-1}S\right)U^\top\| = \|I - (S + \mu I)^{-1}S\|. \end{aligned}$$

This equals one minus the smallest element on the diagonal of the diagonal matrix $(S + \mu I)^{-1}S$, which is $\lambda/(\lambda + \mu)$. □

Lemma 4. *Let A be a positive definite matrix with minimal eigenvalue γ which is larger than some $\mu > 0$, and $\{\Delta_i\}_{i=1}^m$ matrices of the same size, such that $\max_i \|\Delta_i\| \leq \beta$ and $\beta < \gamma$. Then*

$$\left\|\left(\frac{1}{m}\sum_{i=1}^{m}(A + \Delta_i)^{-1} - A^{-1}\right)(A - \mu I)\right\| \leq \frac{2\beta^2}{\gamma\,(\gamma - \beta)}$$

Proof. For any i, we have

$$(A+\Delta_i)^{-1} = \big(A(I+A^{-1}\Delta_i)\big)^{-1} = (I+A^{-1}\Delta_i)^{-1}A^{-1}.$$

Note that $\|A^{-1}\Delta_i\| \leq \|A^{-1}\|\|\Delta_i\| \leq \frac{1}{\gamma}\beta < 1$. Therefore, we can use the identity

$$(I+C)^{-1} = \sum_{r=0}^{\infty}(-1)^r C^r$$

which holds for any C such that $\|C\| < 1$. Using this with $C = A^{-1}\Delta_i$ and plugging back, we get

$$(A+\Delta_i)^{-1} = \sum_{r=0}^{\infty}(-1)^r \left(A^{-1}\Delta_i\right)^r A^{-1} = A^{-1} - A^{-1}\Delta_i A^{-1} + \sum_{r=2}^{\infty}(-1)^r \left(A^{-1}\Delta_i\right)^r A^{-1}.$$

Averaging over $i = 1 \ldots m$ and using the assumption $\sum_{i=1}^m \Delta_i = 0$, we get

$$\frac{1}{m}\sum_{i=1}^{m}(A+\Delta_i)^{-1} = A^{-1} + \sum_{r=2}^{\infty}\frac{(-1)^r}{m}\sum_{i=1}^{m}\left(A^{-1}\Delta_i\right)^r A^{-1}.$$

Multiplying both sides by $(A-\mu I)$, we get

$$\left(\frac{1}{m}\sum_{i=1}^{m}(A+\Delta_i)^{-1}\right)(A-\mu I) = A^{-1}\left(A-\mu I\right) + \sum_{r=2}^{\infty}\frac{(-1)^r}{m}\sum_{i=1}^{m}\left(A^{-1}\Delta_i\right)^r \left(I-\mu A^{-1}\right).$$

By the triangle inequality and convexity of the norm, this implies

$$\begin{aligned}
&\left\|\left(\frac{1}{m}\sum_{i=1}^{m}\left(A+\Delta_i\right)^{-1} - A^{-1}\right)(A-\mu I)\right\| = \left\|\sum_{r=2}^{\infty}\frac{(-1)^r}{m}\sum_{i=1}^{m}\left(A^{-1}\Delta_i\right)^r (I-\mu A^{-1})\right\| \\
&\leq \sum_{r=2}^{\infty}\frac{1}{m}\sum_{i=1}^{m}\|\left(A^{-1}\Delta_i\right)^r (I-\mu A^{-1})\| \\
&\leq \sum_{r=2}^{\infty}\frac{1}{m}\sum_{i=1}^{m}\|A^{-1}\|^r\|\Delta_i\|^r\|I-\mu A^{-1}\| \\
&\leq \sum_{r=2}^{\infty}\frac{\beta^r}{\gamma^r}\left(1+\frac{\mu}{\gamma}\right) \leq \frac{2\beta^2}{\gamma^2}\sum_{r=0}^{\infty}\left(\frac{\beta}{\gamma}\right)^r = \frac{2\beta^2}{\gamma^2}\frac{1}{1-\frac{\beta}{\gamma}} = \frac{2\beta^2}{\gamma(\gamma-\beta)}
\end{aligned}$$

from which the result follows. □

We are now ready to prove Lemma 1. Using Lemma 3, we can upper bound $\|I-\tilde{H}^{-1}H\|$ as

$$\begin{aligned}
\|I - \frac{1}{m}\sum_{i=1}^{m}\left(H_i+\mu I\right)^{-1}H\| &\leq \|I-(H+\mu I)^{-1}H\| + \left\|\frac{1}{m}\sum_{i=1}^{m}\left(H_i+\mu I\right)^{-1}H - (H+\mu I)^{-1}H\right\| \\
&\leq \frac{\mu}{\lambda+\mu} + \left\|\left(\frac{1}{m}\sum_{i=1}^{m}\left(H_i+\mu I\right)^{-1} - \left(H+\mu I\right)^{-1}\right)H\right\|
\end{aligned}$$

Now, we use Lemma 4 with $A = H+\mu I$ and $\Delta_i = H_i - H$ (noting that $\|\Delta_i\| \leq \beta$), and get the bound

$$\frac{\mu}{\lambda+\mu} + \frac{2\beta^2}{\left(\lambda+\mu\right)\left(\lambda+\mu-\beta\right)}.$$

assuming $\beta < \lambda+\mu$.

Now, let us assume the even stronger condition that $\beta < \frac{1}{2}(\lambda+\mu)$ (which we shall justify at the end of the proof), then we can upper bound the right hand side in the equation above by

$$\frac{\mu}{\lambda+\mu} + \frac{4\beta^2}{(\lambda+\mu)^2}. \tag{27}$$

Differentiating with respect to μ, we get an optimal point at

$$\mu^{opt} = \frac{8\beta^2}{\lambda} - \lambda.$$

If this is non-positive, it means that $\lambda^2 > 8\beta^2$, and moreover, that $\mu = 0$, so (27) equals $4\beta^2/\lambda^2$. Otherwise, we pick $\mu = \frac{8\beta^2}{\lambda} - \lambda$, and (27) becomes

$$1 - \frac{\lambda}{\frac{8\beta^2}{\lambda}} + \frac{4\beta^2}{\left(\frac{8\beta^2}{\lambda}\right)^2} = 1 - \frac{\lambda^2}{8\beta^2} + \frac{\lambda^2}{16\beta^2} = 1 - \frac{\lambda^2}{16\beta^2}.$$

Combining the two cases, we get the result stated in the Lemma. Finally, it remains to justify why $\beta < \frac{1}{2}(\lambda+\mu)$. By the way we picked μ, it's enough to prove that

$$2\beta < \max\left\{\lambda, \frac{8\beta^2}{\lambda}\right\},$$

or equivalently,

$$2 < \max\left\{\frac{\lambda}{\beta}, 8\frac{\beta}{\lambda}\right\}.$$

This is true since $\max\{x, 8/x\} > 2$ for all positive x.

D. Proof of Lemma 2

H is the average of the Hessians of mn i.i.d. quadratic functions, all with eigenvalues at most L, and each H_i is the average of the Hessians of n i.i.d. quadratic functions, all with eigenvalues at most L. By a matrix Hoeffding's inequality (Tropp, 2012), we have that for each i, with probability $1-\delta$ over the samples received by machine i,

$$\|H_i - \mathbb{E}[H_i]\| \leq \sqrt{\frac{8L^2\log(d/\delta)}{n}}.$$

By a union bound, we get that with probability $1-\delta$,

$$\max_i \|H_i - \mathbb{E}[H_i]\| \leq \sqrt{\frac{8L^2\log(dm/\delta)}{n}}.$$

Moreover, we have $\mathbb{E}[H_i] = \mathbb{E}[H]$ and $H = \frac{1}{m}\sum_i H_i$, so if this event occurs, we also have

$$\|H - \mathbb{E}[H]\| \leq \sqrt{\frac{8L^2\log(d/\delta)}{n}}.$$

Combining these, we get that with probability $1-\delta$,

$$\max_i \|H_i - H\| \leq \max_i \|H_i - \mathbb{E}[H]\| + \|H - \mathbb{E}[H]\| \leq \sqrt{\frac{32\,L^2\log(dm/\delta)}{n}}.$$

E. Proof of Theorem 3

Plugging 2 into Lemma 1, and noting that the strong convexity of the instantaneous losses implies $\hat{F}(w)$ is λ strongly convex[7], we obtain

$$\|I - \tilde{H}^{-1}H\| \leq \begin{cases} \frac{128(L/\lambda)^2\log(dm/\delta)}{n} & \text{if } \frac{128(L/\lambda)^2\log(dm/\delta)}{n} \leq \frac{1}{2} \\ 1 - \frac{n}{512(L/\lambda)^2\log(dm/\delta)} & \text{otherwise.} \end{cases} \tag{28}$$

[7] In fact, it is enough to require that $\hat{F}(w)$ is λ-strongly convex, and it is not necessary to require strong convexity of $f(w,z)$ for each individual z. However, requiring that the population objective $F(w)$ is λ-strongly convex might not be sufficient if $\lambda < L/\sqrt{n}$, e.g. when $\lambda \propto 1/\sqrt{nm}$.

By smoothness of $\hat{F}$, we have $\hat{F}(w^{(t)}) - \hat{F}(\hat{w}) \leq \frac{L}{2}\|w^{(t)} - \hat{w}\|^2$, and therefore Theorem 2 implies that

$$\hat{F}(w^{(t)}) - \hat{F}(\hat{w}) \leq \frac{L}{2}\|w^{(0)} - \hat{w}\|^2\|I - \eta\tilde{H}^{-1}H\|^{2t}.$$

This means that to get optimization error $\leq \epsilon$, the number of iterations required is

$$\frac{\log\left(\frac{L\|w^{(0)}-\hat{w}\|^2}{2\epsilon}\right)}{-2\log\left(\|I - \eta\tilde{H}^{-1}H\|\right)} \tag{29}$$

Considering (28), if the first case holds, then the denominator in (29) is at least $2\log(2)$ and we get that the number of iterations required is $\mathcal{O}\left(\log\left(\frac{L\|w^{(0)}-\hat{w}\|}{\epsilon}\right)\right)$. If the second case in (28) holds, we have

$$\log\left(\|I - \eta\tilde{H}^{-1}H\|\right) \leq \log\left(1 - \frac{n}{512(L/\lambda)^2\log(dm/\delta)}\right) \leq -\frac{n}{512(L/\lambda)^2\log(dm/\delta)},$$

which implies that the iteration bound (29) is at most

$$\frac{256(L/\lambda)^2\log(dm/\delta)}{n}\log\left(\frac{L\|w^{(0)}-\hat{w}\|^2}{2\epsilon}\right).$$

F. Proof of Theorem 4

We begin with the following lemma:

Lemma 5. *Under the conditions of Theorem 4, the following inequalities hold:*

$$(\nabla h_i(w') - \nabla h_i(w))^\top(w' - w) \geq \frac{1}{L_i + \mu}\|\nabla h_i(w') - \nabla h_i(w)\|_2^2 \tag{30}$$

and

$$\|\nabla\phi(w)\|_2^2 \geq \lambda(\phi(w) - \phi(\hat{w})). \tag{31}$$

Proof. The smoothness of h_i implies that its conjugate h_i^* is $1/(L_i + \mu)$ strongly convex. Let $u' = \nabla h_i(w')$ and $u = \nabla h_i(w)$, then $w' = \nabla h_i^*(u')$ and $w = \nabla h_i^*(u)$. We have

$$(\nabla h_i(w') - \nabla h_i(w))^\top(w' - w) = (\nabla h_i^*(u') - \nabla h_i^*(u))^\top(u' - u) \geq \frac{1}{L_i + \mu}\|u' - u\|_2^2.$$

This proves (30).

Since $\nabla\phi(\hat{w}) = 0$, $\|\nabla\phi(w)\|_2^2 = \|\nabla\phi(w) - \nabla\phi(\hat{w})\|_2^2$. From $\|\nabla\phi(w) - \nabla\phi(\hat{w})\|_2 \geq \lambda\|w - \hat{w}\|_2$, we obtain

$$\begin{aligned}\|\nabla\phi(w) - \nabla\phi(\hat{w})\|_2^2 \geq& \lambda\|\nabla\phi(w) - \nabla\phi(\hat{w})\|_2\|w - \hat{w}\|_2\\ \geq& \lambda(\nabla\phi(w) - \nabla\phi(\hat{w}))^\top(w - \hat{w})\\ =& \lambda\nabla\phi(w)^\top(w - \hat{w}) \geq \lambda(\phi(w) - \phi(\hat{w})).\end{aligned}$$

This proves (31). □

We are now ready to prove the Theorem. At iteration t, we have the following first order equation:

$$\nabla h_i(w_i^{(t)}) - \nabla h_i(w^{(t-1)}) = -\eta\nabla\phi(w^{(t-1)}). \tag{32}$$

Therefore,

$$
\begin{aligned}
&\phi(w_i^{(t)})\\
=&\phi(w^{(t-1)}) + \nabla\phi(w^{(t-1)})^\top(w_i^{(t)} - w^{(t-1)}) + D_\phi(w_i^{(t)}; w^{(t-1)})\\
=&\phi(w^{(t-1)}) - \frac{1}{\eta}(\nabla h_i(w_i^{(t)}) - \nabla h_i(w^{(t-1)}))^\top(w_i^{(t)} - w^{(t-1)}) + D_\phi(w_i^{(t)}; w^{(t-1)})\\
\leq&\phi(w^{(t-1)}) - \frac{1}{\eta}(\nabla h_i(w_i^{(t)}) - \nabla h_i(w^{(t-1)}))^\top(w_i^{(t)} - w^{(t-1)}) + \frac{L}{2}\|w_i^{(t)} - w^{(t-1)}\|_2^2\\
\leq&\phi(w^{(t-1)}) - \frac{1}{\eta}(\nabla h_i(w_i^{(t)}) - \nabla h_i(w^{(t-1)}))^\top(w_i^{(t)} - w^{(t-1)}) + \frac{L}{2(\lambda_i+\mu)^2}\|\nabla h_i(w_i^{(t)}) - \nabla h_i(w^{(t-1)})\|_2^2\\
\leq&\phi(w^{(t-1)}) - \left[\frac{\eta^{-1}}{\mu+L_i} - \frac{L}{2(\lambda_i+\mu)^2}\right]\|\nabla h_i(w_i^{(t)}) - \nabla h_i(w^{(t-1)})\|_2^2\\
=&\phi(w^{(t-1)}) - (\rho_i/\lambda)\|\nabla\phi(w^{(t-1)})\|_2^2.
\end{aligned}
$$

where $\rho_i = \left[\frac{1}{\mu+L_i} - \frac{\eta L}{2(\mu+\lambda_i)^2}\right]\eta\lambda$. In the above derivations, the first inequality uses the smoothness of ϕ; the second inequality uses the strong convexity of h_i; the third inequality uses (30); the second and the last equalities use (32).

Therefore

$$
\begin{aligned}
\phi(w^{(t)}) \leq& \frac{1}{m}\sum_{i=1}^m \phi(w_i^{(t)})\\
\leq& \frac{1}{m}\sum_{i=1}^m [\phi(w^{(t-1)}) - (\rho_i/\lambda)\|\nabla\phi(w^{(t-1)})\|_2^2] = \phi(w^{(t-1)}) - (\rho/\lambda)\|\nabla\phi(w^{(t-1)})\|_2^2\\
\leq& \phi(w^{(t-1)}) - \rho(\phi(w^{(t-1)}) - \phi(\hat{w})),
\end{aligned}
$$

where the first inequality is Jensen's and the third inequality is due to (31). As a result, we get

$$
\phi(w^{(t)}) - \phi(\hat{w}) \leq \phi(w^{(t-1)}) - \phi(\hat{w}) - \rho(\phi(w^{(t-1)}) - \phi(\hat{w})) = (1-\rho)(\phi(w^{(t-1)}) - \phi(\hat{w})).
$$

The desired bound follows by recursively applying the above inequality.

G. Proof of Theorem 5

We have

$$
\begin{aligned}
\phi(\hat{w}) =& \phi(w^{(t-1)}) + \nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t-1)}) + D_\phi(\hat{w}; w^{(t-1)})\\
=&\phi(w^{(t)}) - \nabla\phi(w^{(t-1)})^\top(w^{(t)} - w^{(t-1)}) - D_\phi(w^{(t)}; w^{(t-1)}) + \nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t-1)}) + D_\phi(\hat{w}; w^{(t-1)})\\
=&\phi(w^{(t)}) + \nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t)}) - D_\phi(w^{(t)}; w^{(t-1)}) + D_\phi(\hat{w}; w^{(t-1)})\\
=&\phi(w^{(t)}) + \nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t-1)}) - D_\phi(w^{(t)}; w^{(t-1)}) + D_\phi(\hat{w}; w^{(t-1)}) + \nabla\phi(w^{(t-1)})^\top(w^{(t-1)} - w^{(t)})\\
=&\phi(w^{(t)}) + \nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t-1)}) - D_\phi(w^{(t)}; w^{(t-1)}) + D_\phi(\hat{w}; w^{(t-1)})\\
&+ \eta^{-1}[D_h(w^{(t-1)}; w^{(t)}) + D_h(w^{(t)}; w^{(t-1)})]\\
\geq&\phi(w^{(t)}) + \nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t-1)}) + D_\phi(\hat{w}; w^{(t-1)}) + \eta^{-1}D_h(w^{(t-1)}; w^{(t)}),
\end{aligned}
$$

where in the last inequality, we have used the assumption that $D_\phi(w^{(t)}; w^{(t-1)}) \leq \eta^{-1}D_h(w^{(t)}; w^{(t-1)})$. This implies that

$$
D_h(w^{(t-1)}; w^{(t)}) + \eta\nabla\phi(w^{(t-1)})^\top(\hat{w} - w^{(t-1)}) \leq \eta[\phi(\hat{w}) - \phi(w^{(t)}) - D_\phi(\hat{w}; w^{(t-1)})]. \tag{33}
$$

Therefore we have

$$\begin{aligned}
&D_h(\hat{w}; w^{(t)}) - D_h(\hat{w}; w^{(t-1)}) \\
=&D_h(w^{(t-1)}; w^{(t)}) + (\nabla h(w^{(t-1)}) - \nabla h(w^{(t)}))^\top (\hat{w} - w^{(t-1)}) \\
=&D_h(w^{(t-1)}; w^{(t)}) + \eta \nabla \phi(w^{(t-1)})^\top (\hat{w} - w^{(t-1)}) \\
\leq&\eta[\phi(\hat{w}) - \phi(w^{(t)}) - D_\phi(\hat{w}; w^{(t-1)})] \\
\leq&-\eta D_\phi(\hat{w}; w^{(t-1)}) \\
\leq&-\eta\gamma D_h(\hat{w}; w^{(t-1)}),
\end{aligned}$$

where the first inequality is due to (33), the second inequality comes from the inequality $\phi(\hat{w}) \leq \phi(w^{(t)})$, and the third inequality uses the assumption that $D_\phi(\hat{w}; w^{(t-1)}) \geq \gamma D_h(\hat{w}; w^{(t-1)})$. We thus obtain

$$D_h(\hat{w}; w^{(t)}) \leq (1 - \eta\gamma) D_h(\hat{w}; w^{(t-1)}),$$

and this implies the desired result.