In this appendix, we provide missing proofs in the paper.

A Proof of Proposition 1

We show the convergence of g_λ to the Bayes rule g_* for $\mathcal{L}$ in $\mathcal{H}_k$.

Proposition A. *Let $\mathcal{L}(g)$ be convex with respect to g. Suppose assumption* **(A5)** *holds. A minimizer g_λ of $\mathcal{L}_\lambda$ converges to the Bayes rule g_* in $\mathcal{H}_k$ as $\lambda \to 0$.*

Proof. Let $\{\lambda_i\}_{i=1,2,\dots}$ be a positive decreasing sequence tending to zero in $\mathbb{R}$. Let i, j be arbitrary indices such that $i < j$, i.e., $\lambda_i \geq \lambda_j$. For $g \in \mathcal{H}_k$ satisfying $\|g\|_{\mathcal{H}_k} < \|g_{\lambda_i}\|_{\mathcal{H}_k}$, by subtracting $\frac{\lambda_i - \lambda_j}{2}\|g_{\lambda_i}\|^2_{\mathcal{H}_k} > \frac{\lambda_i - \lambda_j}{2}\|g\|^2_{\mathcal{H}_k}$ from $\mathcal{L}(g) + \frac{\lambda_i}{2}\|g\|^2_{\mathcal{H}_k} \geq \mathcal{L}(g_{\lambda_i}) + \frac{\lambda_i}{2}\|g_{\lambda_i}\|^2_{\mathcal{H}_k}$, we get

$$\mathcal{L}(g) + \frac{\lambda_j}{2}\|g\|^2_{\mathcal{H}_k} > \mathcal{L}(g_{\lambda_i}) + \frac{\lambda_j}{2}\|g_{\lambda_i}\|^2_{\mathcal{H}_k}.$$

This implies that if $\|g\|_{\mathcal{H}_k} < \|g_{\lambda_i}\|_{\mathcal{H}_k}$, then g is not optimal point of $\mathcal{L}_{\lambda_j}$, hence, $\|g_{\lambda_j}\|_{\mathcal{H}_k} \geq \|g_{\lambda_i}\|_{\mathcal{H}_k}$. The boundedness of this sequence is also confirmed because $g_* \in \mathcal{H}_k$ and for $\forall \lambda > 0$,

$$\mathcal{L}(g_*) + \frac{\lambda}{2}\|g_\lambda\|^2_{\mathcal{H}_k} \leq \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|^2_{\mathcal{H}_k} \leq \mathcal{L}(g_*) + \frac{\lambda}{2}\|g_*\|^2_{\mathcal{H}_k}, \tag{1}$$

which implies an inequality $\|g_\lambda\|_{\mathcal{H}_k} \leq \|g_*\|_{\mathcal{H}_k}$. Namely, $\{\|g_{\lambda_i}\|_{\mathcal{H}_k}\}_{i=1,2,\dots}$ is a bounded increasing sequence and has the limit. On the other hand, $\{\mathcal{L}(g_{\lambda_i})\}_{i=1,2,\dots}$ is a decreasing sequence with the limit corresponding to $\mathcal{L}(g_*)$. Indeed, since $\mathcal{L}(g_{\lambda_j}) + \frac{\lambda_j}{2}\|g_{\lambda_j}\|^2_{\mathcal{H}_k} \leq \mathcal{L}(g_{\lambda_i}) + \frac{\lambda_j}{2}\|g_{\lambda_i}\|^2_{\mathcal{H}_k}$, we see

$$0 \leq \frac{\lambda_j}{2}(\|g_{\lambda_j}\|^2_{\mathcal{H}_k} - \|g_{\lambda_i}\|^2_{\mathcal{H}_k}) \leq \mathcal{L}(g_{\lambda_i}) - \mathcal{L}(g_{\lambda_j}).$$

Moreover, from the inequality (1), $\mathcal{L}(g_{\lambda_i})$ converges to $\mathcal{L}(g_*)$.

We next show that the convergence of a sequence $\{g_{\lambda_i}\}_{i=1,2,\dots}$. From the strong convexity of $\mathcal{L}_{\lambda_i}(g)$, we have

$$\mathcal{L}(g_{\lambda_i}) + \frac{\lambda_i}{2}\|g_{\lambda_i}\|^2_{\mathcal{H}_k} + \frac{\lambda_i}{2}\|g_{\lambda_j} - g_{\lambda_i}\|^2_{\mathcal{H}_k} \leq \mathcal{L}(g_{\lambda_j}) + \frac{\lambda_i}{2}\|g_{\lambda_j}\|^2_{\mathcal{H}_k}.$$

Using $\mathcal{L}(g_{\lambda_j}) \leq \mathcal{L}(g_{\lambda_i})$, we get

$$\|g_{\lambda_j} - g_{\lambda_i}\|^2_{\mathcal{H}_k} \leq \|g_{\lambda_j}\|^2_{\mathcal{H}_k} - \|g_{\lambda_i}\|^2_{\mathcal{H}_k} \leq 2\|g_*\|_{\mathcal{H}_k}(\|g_{\lambda_j}\|_{\mathcal{H}_k} - \|g_{\lambda_i}\|_{\mathcal{H}_k}).$$

Since, $\{\|g_{\lambda_i}\|_{\mathcal{H}_k}\}_{i=1,2,\dots}$ is a convergent sequence, it is also a Cauchy sequence. As a result, a sequence $\{g_{\lambda_i}\}_{i=1,2,\dots}$ is Cauchy in $\mathcal{H}_k$ and has a limit point $g_\infty \in \mathcal{H}_k$. It follows from the continuity of $\mathcal{L}$ that $\mathcal{L}(g_\infty) = \lim_{i\to\infty}\mathcal{L}(g_{\lambda_i})$. Recalling $\lim_{i\to\infty}\mathcal{L}(g_{\lambda_i}) = \mathcal{L}(g_*)$ and the uniqueness of the Bayes rule g_*, we conclude $g_\infty = g_*$ up to zero measure sets. □

We now give a proof of Proposition 1.

Proof of Proposition 1 . Noting that $g(x) = \langle g, k(x,\cdot)\rangle_{\mathcal{H}_k}$ for arbitrary function $g \in \mathcal{H}_k$ and $k(x,\cdot) \in \mathcal{H}_k$ by the definition of kernel function, we get

$$\|g\|_{L_\infty} = \sup_{x\in\mathcal{X}}|g(x)| \leq \|g\|_{\mathcal{H}_k}\|k(x,\cdot)\|_{\mathcal{H}_k} \leq R\|g\|_{\mathcal{H}_k}. \tag{2}$$

Since, g_λ converges to g_* in $\mathcal{H}_k$ from Proposition A, there exists $\lambda > 0$ such that

$$\|g_\lambda - g_*\|_{\mathcal{H}_k} \leq \frac{m(\delta)}{2R}.$$

Thus, for arbitrary $g \in \mathcal{H}_k$ satisfying $\|g - g_\lambda\|_{\mathcal{H}_k} \leq \frac{m(\delta)}{2R}$, we have

$$\|g - g_*\|_{L_\infty} \leq R\|g - g_*\|_{\mathcal{H}_k} \leq R\left(\|g - g_\lambda\|_{\mathcal{H}_k} + \|g_\lambda - g_*\|_{\mathcal{H}_k}\right) \leq m(\delta).$$

Since, $m(\delta) \leq |g_*(X)|$ almost surely, we get $\mathrm{sgn}(g_*(X)) = \mathrm{sgn}(g(X))$ almost surely for $g \in \mathcal{H}_k$ such that $\|g - g_\lambda\|_{\mathcal{H}} \leq \frac{m(\delta)}{2R}$, that is, g is also the Bayes rule for $\mathcal{R}$. □

B Proof of Theorem 1

In this section, we give proofs of auxiliary statements needed for the main theorem meaning the exponential convergence of stochastic gradient descent. We here prove convergence of expected functions obtained by stochastic gradient descent.

Proposition 3. By $(L+\lambda)$-Lipschitz smoothness $\mathcal{L}_\lambda$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_\lambda(g_t - \eta_t G_\lambda(g_t, z_t))] &\leq \mathbb{E}[\mathcal{L}_\lambda(g_t)] - \eta_t \mathbb{E}[\langle \nabla \mathcal{L}_\lambda(g_t), G_\lambda(g_t, z_t)\rangle_{\mathcal{H}_k}] + \frac{(L+\lambda)\eta_t^2}{2}\mathbb{E}\|G_\lambda(g_t, z_t)\|_{\mathcal{H}_k}^2 \\ &\leq \mathbb{E}[\mathcal{L}_\lambda(g_t)] - \eta_t \mathbb{E}\|\nabla \mathcal{L}_\lambda(g_t)\|_{\mathcal{H}_k}^2 + \frac{(L+\lambda)\eta_t^2}{2}(\mathbb{E}\|\nabla \mathcal{L}_\lambda(g_t)\|_{\mathcal{H}_k}^2 + \sigma^2) \\ &\leq \mathbb{E}[\mathcal{L}_\lambda(g_t)] - \frac{\eta_t}{2}\mathbb{E}\|\nabla \mathcal{L}_\lambda(g_t)\|_{\mathcal{H}_k}^2 + \frac{(L+\lambda)\eta_t^2\sigma^2}{2}, \end{aligned} \tag{3}$$

where we used $\eta_t \leq 1/(L+\lambda)$ for the last inequality. On the other hand, by the strong convexity of $\mathcal{L}_\lambda$, we have for $\forall g \in \mathcal{H}_k$,

$$\mathcal{L}_\lambda(g_t) + \langle \nabla \mathcal{L}_\lambda(g_t), g - g_t\rangle_{\mathcal{H}_k} + \frac{\lambda}{2}\|g - g_t\|_{\mathcal{H}_k}^2 \leq \mathcal{L}_\lambda(g).$$

Minimizing both sides with respect to g in $\mathcal{H}_k$, we have

$$\mathcal{L}_\lambda(g_t) - \frac{1}{2\lambda}\|\nabla \mathcal{L}_\lambda(g_t)\|_{\mathcal{H}_k}^2 \leq \mathcal{L}_\lambda(g_\lambda). \tag{4}$$

By combining two inequalities (3) and (4) and subtracting $\mathcal{L}_\lambda(g_\lambda)$, we get

$$\mathbb{E}[\mathcal{L}_\lambda(g_{t+1})] - \mathcal{L}_\lambda(g_\lambda) \leq (1 - \eta_t\lambda)\left(\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda)\right) + \frac{(L+\lambda)\eta_t^2\sigma^2}{2}. \tag{5}$$

We now show the following convergence rate by induction on t.

$$\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda) \leq \frac{\nu}{\gamma + t}. \tag{6}$$

For $t = 1$, it is clearly true from the choice of ν. We suppose that the inequality (6) is true for t. We denote $\hat{t} = \gamma + t$ for simplicity. Then, we have that from the inequality (5) and $\eta_t = 2/\lambda\hat{t}$,

$$\begin{aligned}\mathbb{E}[\mathcal{L}_\lambda(g_{t+1})] - \mathcal{L}_\lambda(g_\lambda) &\leq \left(1 - \frac{2}{\hat{t}}\right)\frac{\nu}{\hat{t}} + \frac{2(L+\lambda)\sigma^2}{\lambda^2\hat{t}^2} \\ &= \frac{(\hat{t}-1)\nu}{\hat{t}^2} - \frac{\nu}{\hat{t}^2} + \frac{2(L+\lambda)\sigma^2}{\lambda^2\hat{t}^2} \\ &\leq \frac{\nu}{\hat{t}+1},\end{aligned}$$

where we used $\hat{t}^2 > (\hat{t}+1)(\hat{t}-1)$ and the definition of ν. Thus, the inequality (6) is true for all $T \geq 1$. From the strong convexity and Jensen's inequality for $\mathcal{L}_\lambda$, we have

$$\|\mathbb{E}[g_t] - g_\lambda\|_{\mathcal{H}_k}^2 \leq \frac{2}{\lambda}(\mathcal{L}_\lambda(\mathbb{E}[g_t]) - \mathcal{L}_\lambda(g_\lambda)) \leq \frac{2}{\lambda}(\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda)).$$

This finishes the proof of the proposition. □

As argued in the paper, the proof of Proposition 4 is reduced to bounding $\|g_{T+1} - g_{T+1}^t\|_\infty$. The following proposition is useful for that purpose. Let g_s^t ($s \geq t \in \{1, \ldots, T+1\}$) be the s-th iterate depending on $(Z_1, \ldots, Z_{t-1}, Z_t', Z_{t+1}, \ldots, Z_s)$.

Proposition B. *Suppose Assumptions* **(A1)** *and* **(A2)** *hold. We consider Algorithm 1 without the averaging option and with a decreasing learning rates η_t. We assume that $\|g_1\|_{\mathcal{H}_k} \leq (2\eta_1 + 1/\lambda)MR$ and $\eta_1 \leq \min\{1/L, 1/2\lambda\}$. Then, for $t \in \{1, \ldots, T\}$, it follows that*

1. $\|g_{t+1} - g_{t+1}^t\|_{\mathcal{H}_k} \leq 6MR\eta_t$,
2. $\|g_{s+1} - g_{s+1}^t\|_{\mathcal{H}_k} \leq (1 - \eta_s\lambda)\|g_s - g_s^t\|_{\mathcal{H}_k}$ *for* $s \geq t+1$.

Proof. By the assumptions, we find that the stochastic gradient of l in $\mathcal{H}_k$ is bounded as follows:

$$\|\partial_\varsigma l(g(x), y)k(x, \cdot)\|_{\mathcal{H}_k} \leq MR.$$

Therefore, if $\|g_t\|_{\mathcal{H}_k} \geq \frac{1}{\lambda}MR$, then

$$\begin{aligned}\|g_{t+1}\|_{\mathcal{H}_k} &= \|g_t - \eta_t\partial_\varsigma l(g(X_t), Y_t)k(X_t, \cdot) - \eta_t\lambda g_t\|_{\mathcal{H}} \\ &\leq (1 - \eta_t\lambda)\|g_t\|_{\mathcal{H}_k} + \eta_t MR \\ &\leq \|g_t\|_{\mathcal{H}_k}.\end{aligned}$$

This means a generated sequence $\{g_t\}_{t=1,\ldots,T+1}$ is included in a closed ball centered at the origin with radius $(2\eta_1 + 1/\lambda)MR$ as long as an initial function g_1 is contained in this ball. Thus, the norm of $G_\lambda(g_t, Z_t)$ is bounded by $2(1 + \lambda\eta_1)MR \leq 3MR$.

The first statement can be shown as follows: since $g_t = g_t^t$,

$$\|g_{t+1} - g_{t+1}^t\|_{\mathcal{H}_k} = \eta_t\|G_\lambda(g_t, Z_t) - G_\lambda(g_t, Z_t')\|_{\mathcal{H}_k} \leq 6\eta_t MR.$$

We next show the second statement. The Lipschitz smoothness of $\mathcal{L}$ leads to the following inequality which can be confirmed by naturally extending the proof of [Nes04] to the Hilbert space. Let $\partial_g l(g, z)$ denote the gradient of $l(g, z)$ with respect to g in $\mathcal{H}_k$. Then, we have for $\forall g, \forall g' \in \mathcal{H}_k$,

$$\langle \partial_g l(g, z) - \partial_g l(g', z), g - g' \rangle_{\mathcal{H}_k} \geq \frac{1}{L}\|\partial_g l(g, z) - \partial_g l(g', z)\|_{\mathcal{H}_k}^2. \tag{7}$$

Thus, we have that for $s \geq t+1$,

$$\begin{aligned}\|g_{s+1} - g_{s+1}^t\|_{\mathcal{H}_k}^2 &= \|(1 - \eta_s\lambda)(g_s - g_s^t) - \eta_s(\partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s))\|_{\mathcal{H}_k}^2 \\ &= (1 - \eta_s\lambda)^2\|g_s - g_s^t\|_{\mathcal{H}_k}^2 - 2\eta_s(1 - \eta_s\lambda)\left\langle \partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s), g_s - g_s^t \right\rangle \\ &+ \eta_s^2\|\partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s)\|_{\mathcal{H}_k}^2 \\ &\leq (1 - \eta_s\lambda)^2\|g_s - g_s^t\|_{\mathcal{H}_k}^2 - \eta_s\left(\frac{1}{L} - \eta_s\right)\|\partial_g l(g_s, Z_s) - \partial_g l(g_s^t, Z_s)\|_{\mathcal{H}_k}^2 \\ &\leq (1 - \eta_s\lambda)^2\|g_s - g_s^t\|_{\mathcal{H}_k}^2,\end{aligned}$$

where we used the inequality (7) and conditions on learning rates. □

Utilizing this proposition, the stable property of stochastic gradient descent is shown.

Proof of Proposition 4. From Proposition B, we immediately obtain the bound: for $t \in \{1, \ldots, T\}$,

$$\|g_{T+1} - g_{T+1}^t\|_{\mathcal{H}_k} \leq 6MR\eta_t \prod_{s=t+1}^{T}(1 - \eta_s\lambda). \tag{8}$$

From the following inequality,

$$\prod_{s=2}^{T}(1 - \eta_s\lambda) = \prod_{s=2}^{T}\frac{\gamma + s - 2}{\gamma + s} < \frac{\gamma}{\gamma + T},$$

where the last inequality hold clearly by expanding the product, the right hand side of the inequality (8) is upper bounded as follows

$$6MR\eta_t \prod_{s=t+1}^{T}(1 - \eta_s\lambda) \leq 6MR\eta_t\frac{\gamma}{\gamma + T}\frac{\gamma + t}{\gamma} = \frac{12MR}{\lambda(\gamma + T)}.$$

We finally obtain the desired bound:

$$\sum_{t=1}^{T}\|D_t\|_\infty^2 \leq \sum_{t=1}^{T}\frac{144M^2R^2}{\lambda^2(\gamma + T)^2} \leq \frac{144M^2R^2}{\lambda^2(\gamma + T)}.$$

□

C Proof of Theorem 2

In this section we provide auxiliary results for showing Theorem 2. Using them, we can show the theorem in the same way as in the case of stochastic gradient descent without averaging.

We first give a convergence rate of expected functions obtained by averaged stochastic gradient descent. Recall that $\overline{g}_{T+1} = \sum_{t=1}^{T+1} \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)} g_t$.

Proposition C. *Let the loss function ϕ be convex, that is, let $l(g(x), y)$ be also convex with respect to g. Consider Algorithm 1 with the averaging option. Learning rates and averaging weights are $\eta_t = 2/\lambda(\gamma + t)$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$, respectively. Then, it follows that*

$$\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\|_{\mathcal{H}_k}^2 \leq \frac{2}{\lambda}\left(\frac{18M^2R^2}{\lambda(2\gamma+T)} + \frac{\lambda\gamma(\gamma-1)}{2(2\gamma+T)(T+1)}\|g_1 - g_\lambda\|_{\mathcal{H}_k}^2\right).$$

Proof. Recall that the norm of the stochastic gradient $G_\lambda(g_t, Z_t)$ can be upper-bounded by $3MR$ as shown in the proof of Proposition B. Combining this with the strong convexity of $\mathcal{L}_\lambda$, we have

$$\begin{aligned}
\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}_k}^2 &= \mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2 - 2\eta_t\mathbb{E}[\langle g_t - g_\lambda, G_\lambda(g_t, Z_t)\rangle_{\mathcal{H}_k}] + \eta_t^2\mathbb{E}\|G_\lambda(g_t, Z_t)\|_{\mathcal{H}_k}^2 \\
&\leq \mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2 - 2\eta_t\mathbb{E}[\langle g_t - g_\lambda, \nabla\mathcal{L}_\lambda(g_t)\rangle_{\mathcal{H}_k}] + 9\eta_t^2M^2R^2 \\
&\leq \mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2 - 2\eta_t\left(\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda) + \frac{\lambda}{2}\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2\right) + 9\eta_t^2M^2R^2
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda) &\leq \frac{9\eta_tM^2R^2}{2} + \frac{1-\lambda\eta_t}{2\eta_t}\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2 - \frac{1}{2\eta_t}\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}_k}^2 \\
&= \frac{9M^2R^2}{\lambda(\gamma+t)} + \frac{\lambda(\gamma+t-2)}{4}\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2 - \frac{\lambda(\gamma+t)}{4}\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}_k}^2.
\end{aligned}$$

By multiplying $\gamma + t - 1$ and taking sum over $t \in \{1, \ldots, T+1\}$, we get

$$\begin{aligned}
\sum_{t=1}^{T+1}(\gamma+t-1)(\mathbb{E}[\mathcal{L}_\lambda(g_t)] - \mathcal{L}_\lambda(g_\lambda)) &< \frac{9M^2R^2T}{\lambda} + \frac{\lambda}{4}\sum_{t=1}^{T+1}\{(\gamma+t-1)(\gamma+t-2))\mathbb{E}\|g_t - g_\lambda\|_{\mathcal{H}_k}^2 \\
&\quad - (\gamma+t)(\gamma+t-1)\mathbb{E}\|g_{t+1} - g_\lambda\|_{\mathcal{H}_k}^2\} \\
&\leq \frac{9M^2R^2(T+1)}{\lambda} + \frac{\lambda}{4}\gamma(\gamma-1)\|g_1 - g_\lambda\|_{\mathcal{H}_k}^2.
\end{aligned}$$

Thus, by dividing $(2\gamma + T)(T + 1)/2$ and applying Jensen's inequality for $\mathcal{L}_\lambda$, the following convergence rate is obtained:

$$\mathbb{E}\left[\mathcal{L}_\lambda\left(\sum_{t=1}^{T+1}\frac{2(\gamma+t-1)g_t}{(2\gamma+T)(T+1)}\right) - \mathcal{L}_\lambda(g_\lambda)\right] \leq \frac{18M^2R^2}{\lambda(2\gamma+T)} + \frac{\lambda\gamma(\gamma-1)}{2(2\gamma+T)(T+1)}\|g_1 - g_\lambda\|_{\mathcal{H}_k}^2.$$

Thus, the desired inequality is obtained by Jensen's inequality and the strong convexity of $\mathcal{L}_\lambda$. □

Proposition D. *Suppose the same assumptions as in Proposition B. Consider Algorithm 1 with the averaging option. Learning rates and averaging weights are $\eta_t = 2/\lambda(\gamma + t)$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$, respectively. Then, it follows that*

$$\sum_{t=1}^{T}\|D_t\|_\infty^2 \leq \frac{288M^2R^2}{\lambda^2(2\gamma+T)},$$

where $D_t = \mathbb{E}[\overline{g}_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[\overline{g}_{T+1}|Z_1, \ldots, Z_{t-1}]$.

Proof. Note that $\overline{g}_{t+1} \leftarrow (1-\beta_t)\overline{g}_t + \beta_t g_{t+1}$, where $\beta_t = \frac{2(\gamma+t)}{(t+1)(2\gamma+t)}$. Thus, we have

$$\|\overline{g}_{T+1} - \overline{g}^t_{T+1}\|_{\mathcal{H}_k} \leq (1-\beta_T)\|\overline{g}_T - \overline{g}^t_T\|_{\mathcal{H}_k} + \beta_T \|g_{T+1} - g^t_{T+1}\|_{\mathcal{H}_k}.$$

By recursively expanding updates, we obtain the following upper-bound:

$$\sum_{s=t}^{T} \left\{ \prod_{r=s+1}^{T} (1-\beta_r) \right\} \beta_s \|g_{s+1} - g^t_{s+1}\|_{\mathcal{H}_k}.$$

Recall the proof of Proposition 4, it follows that

$$\|g_{s+1} - g^t_{s+1}\|_{\mathcal{H}_k} \leq 6MR\eta_t \prod_{r=t+1}^{s} (1-\eta_r\lambda) \leq \frac{12MR}{\lambda(\gamma+s)}.$$

Since $\prod_{r=s+1}^{T}(1-\beta_r) = \frac{(s+1)(2\gamma+s)}{(T+1)(2\gamma+T)}$, we have

$$\|\overline{g}_{T+1} - \overline{g}^t_{T+1}\|_{\mathcal{H}_k} \leq \sum_{s=t}^{T} \frac{24MR}{\lambda(T+1)(2\gamma+T)} = \frac{24MR(T-t+1)}{\lambda(T+1)(2\gamma+T)}.$$

Therefore, we have the following bound: since $\sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6$,

$$\begin{aligned}
\sum_{t=1}^{T} \|D_t\|^2_\infty &\leq \frac{24^2 M^2 R^2}{\lambda^2 (T+1)^2 (2\gamma+T)^2} \sum_{t=1}^{T} (T-t+1)^2 \\
&\leq \frac{24 \cdot 4M^2R^2(2T+1)}{\lambda^2(2\gamma+T)^2} \\
&\leq \frac{288M^2R^2}{\lambda^2(2\gamma+T)}.
\end{aligned}$$

□

References

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.