B SUPPLEMENTARY MATERIAL

B.1 PROOF FOR THEOREM 4

Notation. For a positive definite matrix $H \in \mathbb{R}^{d\times d}$, the weighted ℓ_2-norm is defined by $\|\mathbf{x}\|_H^2 = \mathbf{x}^\top H\mathbf{x}$. The H-weighted projection $P_{\mathbf{Q}}^H(\mathbf{x})$ of $\mathbf{x}$ onto $\mathbf{Q}$ is defined by $P_{\mathbf{Q}}^H(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbf{Q}} \|\mathbf{y} - \mathbf{x}\|_H^2$. We use $\mathbf{g}(\mathbf{w}_k)$ to denote the subgradient of $f_k(\cdot)$ at $\mathbf{w}_k$. For the diagonal matrix sequence $\{M_k\}_{k=1}^t$, we use $m_{k,i}$ to denote the i-th element in the diagonal of M_k. We use $g_{k,i}$ to denote the i-th element of $\mathbf{g}(\mathbf{w}_k)$.

Lemma 7. *(Mukkamala & Hein, 2017) Suppose that* $1 - \frac{1}{t} \le \beta_{2t} \le 1 - \frac{\gamma}{t}$ *for some* $0<\gamma\le 1$*, and* $t \ge 1$*, then*

$$\sum_{i=1}^{d}\sum_{k=1}^{t} \frac{g_{k,i}^2}{\sqrt{k v_{k,i}} + \delta} \le \sum_{i=1}^{d} \frac{2(2-\gamma)}{\gamma}(\sqrt{t v_{t,i}} + \delta).$$

Proof for Theorem 4. Without loss of generality, we only prove Theorem 4 in the full gradient setting. It can be extended to stochastic cases using the regular technique in (Rakhlin et al., 2011).

Note that the projection operation can be rewritten as an optimization problem (Duchi, 2018), i.e., $\mathbf{w}_{t+1} = P_{\mathbf{Q}}[\mathbf{w}_t - \alpha_t \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t) + \beta_{1t}(\mathbf{w}_t - \mathbf{w}_{t-1})]$ is equivalent to

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbf{Q}} \{\alpha_t \hat{V}_t^{-1}\langle \mathbf{g}(\mathbf{w}_t), \mathbf{w}\rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t - \beta_{1t}(\mathbf{w}_t - \mathbf{w}_{t-1})\|^2\}. \tag{14}$$

Then, $\forall \mathbf{u} \in \mathbf{Q}$, we have

$$\langle \mathbf{w}_{t+1} - \mathbf{w}_t - \beta_t(\mathbf{w}_t - \mathbf{w}_{t-1}) + \alpha_t \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}\rangle \le 0.$$

This is

$$\langle \mathbf{w}_{t+1} + \mathbf{p}_{t+1} - (\mathbf{w}_t + \mathbf{p}_t) + \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}\rangle \le 0. \tag{15}$$

Specifically,

$$\langle \mathbf{w}_{t+1} + \mathbf{p}_{t+1} - (\mathbf{w}_t + \mathbf{p}_t) + \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle \le 0. \tag{16}$$

From (15) and (16),

$$\langle \mathbf{w}_{t+1} + \mathbf{p}_{t+1} - (\mathbf{w}_t + \mathbf{p}_t) + \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t + (t+1)(\mathbf{w}_{t+1} - \mathbf{w}_t)\rangle \le 0.$$

i.e.,

$$\langle \mathbf{w}_{t+1} + \mathbf{p}_{t+1} - (\mathbf{w}_t + \mathbf{p}_t) + \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t+1} + \mathbf{p}_{t+1} - \mathbf{w}_t\rangle \le 0.$$

Using Lemma 6, we have

$$\mathbf{w}_{t+1} + \mathbf{p}_{t+1} = P_{\mathbf{Q}}^{\hat{V}_t}[\mathbf{w}_t + \mathbf{p}_t - \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t)].$$

Then

$$\|\mathbf{w}^* - (\mathbf{w}_{t+1} + \mathbf{p}_{t+1})\|_{\hat{V}_t}^2 \le \|\mathbf{w}^* - (\mathbf{w}_t + \mathbf{p}_t) + \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t)\|_{\hat{V}_t}^2.$$

$$\|\mathbf{w}^* - (\mathbf{w}_t + \mathbf{p}_t) + \frac{\alpha}{\sqrt{t}} \hat{V}_t^{-1}\mathbf{g}(\mathbf{w}_t)\|_{\hat{V}_t}^2$$

$$= \|\mathbf{w}^* - (\mathbf{w}_t + \mathbf{p}_t)\|_{\hat{V}_t}^2 + \|\frac{\alpha}{\sqrt{t}}\mathbf{g}(\mathbf{w}_t)\|_{\hat{V}_t}^2 + 2\langle \frac{\alpha}{\sqrt{t}}\mathbf{g}(\mathbf{w}_t), \mathbf{w}^* - \mathbf{w}_t\rangle + 2\langle \frac{\alpha t}{\sqrt{t}}\mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t-1} - \mathbf{w}_t\rangle.$$

Note

$$\langle \mathbf{g}(\mathbf{w}_t), \mathbf{w}^* - \mathbf{w}_t\rangle \le f(\mathbf{w}^*) - f(\mathbf{w}_t),\ \langle \mathbf{g}(\mathbf{w}_t), \mathbf{w}_{t-1} - \mathbf{w}_t\rangle \le f(\mathbf{w}_{t-1}) - f(\mathbf{w}_t).$$

Then

$$(t+1)(f(\mathbf{w}_t) - f(\mathbf{w}^*))$$

$$\le t(f(\mathbf{w}_{t-1}) - f(\mathbf{w}^*)) + \frac{\sqrt{t}}{2\alpha}\|\mathbf{w}^* - (\mathbf{w}_t + \mathbf{p}_t)\|_{\hat{V}_t}^2 - \frac{\sqrt{t}}{2\alpha}\|\mathbf{w}^* - (\mathbf{w}_{t+1} + \mathbf{p}_{t+1})\|_{\hat{V}_t}^2 + \frac{\alpha}{2\sqrt{t}}\|\mathbf{g}(\mathbf{w}_t)\|_{\hat{V}_t^{-1}}^2.$$

Summing this inequality from $k=1$ to t, we obtain

$$
\begin{aligned}
&(t+1)(f(\mathbf{w}_t)-f(\mathbf{w}^*)) \\
&\leq f(\mathbf{w}_0)-f(\mathbf{w}^*)+\sum_{k=1}^{t}\frac{\alpha}{2\sqrt{k}}\|\mathbf{g}(\mathbf{w}_k)\|^2_{\hat{V}_k^{-1}}+\sum_{k=1}^{t}\big[\frac{\sqrt{k}}{2\alpha}(\|\mathbf{w}^*-(\mathbf{w}_k+\mathbf{p}_k)\|^2_{\hat{V}_k}-\|\mathbf{w}^*-(\mathbf{w}_{k+1}+\mathbf{p}_{k+1})\|^2_{\hat{V}_k})\big].
\end{aligned}
$$

Using Lemma 7, we have

$$
\sum_{k=1}^{t}\frac{\alpha}{2\sqrt{k}}\|\mathbf{g}(\mathbf{w}_k)\|^2_{\hat{V}_k^{-1}}\leq\sum_{i=1}^{d}\frac{\alpha(2-\gamma)}{\gamma}(\sqrt{tv_{t,i}}+\delta).
$$

Note

$$
\sqrt{v_{t,i}}\leq M.
$$

and

$$
\begin{aligned}
&\sum_{k=1}^{t}\big[\frac{\sqrt{k}}{2\alpha}(\|\mathbf{w}^*-(\mathbf{w}_k+\mathbf{p}_k)\|^2_{\hat{V}_k}-\|\mathbf{w}^*-(\mathbf{w}_{k+1}+\mathbf{p}_{k+1})\|^2_{\hat{V}_k})\big] \\
&=\sum_{i=1}^{d}\frac{\hat{V}_1}{2\alpha}\|\mathbf{w}^*-(\mathbf{w}_1+\mathbf{p}_1)\|^2-\sum_{i=1}^{d}\frac{\sqrt{t}\hat{V}_t}{2\alpha}\|\mathbf{w}^*-(\mathbf{w}_{t+1}+\mathbf{p}_{t+1})\|^2 \\
&+\sum_{i=1}^{d}\sum_{k=2}^{t}\frac{1}{2\alpha}(\sqrt{k}\hat{v}_{k,i}-\sqrt{k-1}\hat{v}_{k-1,i})\|\mathbf{w}^*-(\mathbf{w}_k+\mathbf{p}_k)\|^2.
\end{aligned}
\tag{17}
$$

Since $\mathbf{Q}$ is a bounded set, there exists a positive number $M_0>0$ such that

$$
\|\mathbf{w}^*-(\mathbf{w}_{t+1}+\mathbf{p}_{t+1})\|^2\leq M_0,\forall t\geq 0.
$$

and $v_{k,i}=\beta_{2k}v_{k-1,i}+(1-\beta_{2k})g^2_{k,i}$ as well as $\beta_{2k}\geq 1-\frac{1}{k}$ which implies $k\beta_{2k}\geq k-1$, we get

$$
\begin{aligned}
\sqrt{k}\hat{v}_{k,i}&=\sqrt{kv_{k,i}}+\delta \\
&=\sqrt{k\beta_{2k}v_{k-1,i}+k(1-\beta_{2k})g^2_{k,i}}+\delta \\
&\geq\sqrt{(k-1)v_{k-1,i}}+\delta \\
&=\sqrt{k-1}\hat{v}_{k-1,i}.
\end{aligned}
$$

From (17) we have

$$
\begin{aligned}
&\sum_{k=1}^{t}\big[\frac{\sqrt{k}}{2\alpha}(\|\mathbf{w}^*-(\mathbf{w}_k+\mathbf{p}_k)\|^2_{\hat{V}_k}-\|\mathbf{w}^*-(\mathbf{w}_{k+1}+\mathbf{p}_{k+1})\|^2_{\hat{V}_k})\big] \\
&\leq\sum_{i=1}^{d}\frac{\hat{v}_{1,i}}{2\alpha}M_0+\sum_{i=1}^{d}\sum_{k=2}^{t}\frac{1}{2\alpha}(\sqrt{k}\hat{v}_{k,i}-\sqrt{k-1}\hat{v}_{k-1,i})M_0 \\
&=\frac{d\hat{v}_{1,i}M_0}{2\alpha}+\frac{d\sqrt{t}\hat{v}_{t,i}M_0}{2\alpha}-\frac{d\hat{v}_{1,i}M_0}{2\alpha} \\
&\leq\frac{d(\sqrt{t}M+\delta)M_0}{2\alpha}.
\end{aligned}
\tag{18}
$$

Therefore

$$
(t+1)[f(\mathbf{w}_t)-f(\mathbf{w}^*)]\leq f(\mathbf{w}_0)-f(\mathbf{w}^*)+\frac{d\alpha(2-\gamma)(\sqrt{t}M+\delta)}{\gamma}+\frac{d(\sqrt{t}M+\delta)M_0}{2\alpha}.
$$

This proves

$$
f(\mathbf{w}_t)-f(\mathbf{w}^*)\leq O(\frac{1}{\sqrt{t}}).
$$

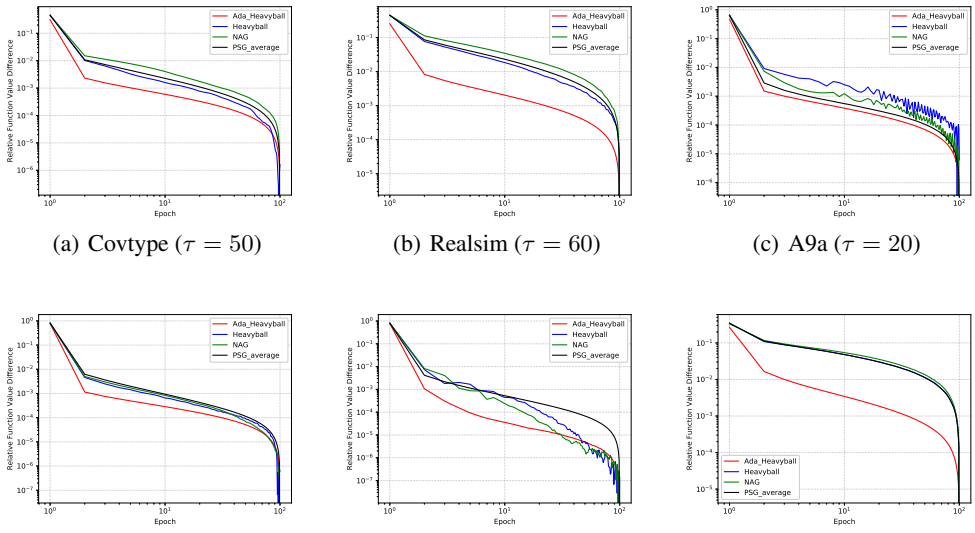
B.2 EXPERIMENTS ON OPTIMIZING GENERAL CONVEX FUNCTIONS

We consider the hinge loss optimization problem with l_1-ball constraints and use SLEP package[1] for l_1 projection operation.

$$\min f(\mathbf{w}),\ s.t.\ \mathbf{w} \in \{\mathbf{w} : \|\mathbf{w}\|_1 \leq \tau\}. \tag{19}$$

Datasets: A9a, W8a, Covtype, Ijcnn1, Rcv1, Realsim (available at LibSVM[2] website).

Algorithms: PSG ($\alpha_t \equiv \frac{\alpha}{\sqrt{t}}$), HB ($\alpha_t \equiv \frac{\alpha}{\sqrt{t}}$, $\beta_t \equiv 0.9$), NAG (Tao et al., 2020a) and adaptive HB (8) ($\beta_{1t} = \frac{t}{t+2}$).



(a) Covtype ($\tau = 50$) (b) Realsim ($\tau = 60$) (c) A9a ($\tau = 20$)

(d) W8a ($\tau = 30$) (e) Ijcnn1 ($\tau = 10$) (f) Rcv1 ($\tau = 80$)

Figure 5: Convergence on different LibSVM datasets for l_1 hinge loss problems

The relative function value $f(\mathbf{w}_t) - f(\mathbf{w}_*)$ v.s. epoch is illustrated in Figure 5. As expected, the individual convergence of the adaptive HB has almost the same behavior as the averaging output of PSG, and the individual output of HB and NAG. Since the three stochastic methods have the optimal convergence, we conclude that the stochastic adaptive HB attains the optimal individual convergence for general convex regularized learning problems.

[1] http://yelabs.net/software/SLEP/

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/