A Proofs

We give proofs of convergence analyses. We first prove the Proposition 1.

Proof of Proposition 1. Since ϕ_k is $(L_g + \mu_k)$-smooth function, we have

$$\phi_k(x) \le \phi_k(x_k) + \langle \nabla \phi_k(x_k), x - x_k \rangle + \frac{L_g + \mu_k}{2} \|x - x_k\|_2^2.$$

By minimizing both sides of the above inequality,

$$\phi_k^* \le \phi_k(x_k) - \frac{1}{2(L_g + \mu_k)} \|\nabla \phi_k(x_k)\|_2^2.$$

Noting that $\phi_k(x_k) = f(x_k)$ and $\mathbb{E}_{v_h(x_k)}[\|\nabla \phi_k(x_k)\|_2^2 | \mathcal{F}_k] \ge \|\nabla f(x_k)\|_2^2$, we have

$$\mathbb{E}_{v_h(x_k)}[\phi_k^* | \mathcal{F}_k] \le f(x_k) - \frac{1}{2(L_g + \mu_k)} \|\nabla f(x_k)\|_2^2.$$

Using $\mathbb{E}[\phi_k(x_{k+1}) | \mathcal{F}_k] \le \phi_k^* + \delta$ and the above inequality, we have

$$\mathbb{E}[\phi_k(x_{k+1}) | \mathcal{F}_k] \le \delta + f(x_k) - \frac{1}{2(L_g + \mu_k)} \|\nabla f(x_k)\|_2^2.$$

Thus, it follows that

$$\begin{aligned}
&\mathbb{E}[f(x_{k+1}) + \frac{\mu_k}{2} \|x_{k+1} - x_k\|_2^2 | \mathcal{F}_k] \\
&\le \mathbb{E}[g(x_{k+1}) - (h(x_k) + \langle \nabla h(x_k), x_{k+1} - x_k \rangle) + \frac{\mu_k}{2} \|x_{k+1} - x_k\|_2^2 | \mathcal{F}_k] \\
&= \mathbb{E}[\phi_k(x_{k+1}) - \langle \nabla h(x_k) - v_h(x_k), x_{k+1} - x_k \rangle | \mathcal{F}_k] \\
&\le \mathbb{E}[\phi_k(x_{k+1}) | \mathcal{F}_k] + \mathbb{E}[\frac{1}{\mu_k} \|\nabla h(x_k) - v_h(x_k)\|_2^2 + \frac{\mu_k}{4} \|x_{k+1} - x_k\|_2^2 | \mathcal{F}_k] \\
&\le \delta + f(x_k) - \frac{1}{2(L_g + \mu_k)} \|\nabla f(x_k)\|_2^2 + \frac{\sigma_h^2}{\mu_k} + \frac{\mu_k}{4} \mathbb{E}[\|x_{k+1} - x_k\|_2^2 | \mathcal{F}_k],
\end{aligned}$$

where for the first inequality we used convexity of h and for the second inequality we used Young's inequality. This finishes the proof of Proposition 1. □

Next, let us prove Theorem 1.

Proof of Theorem 1. Summing up the inequality of Proposition 1 over indices $k = 1, \ldots, M$ and taking the expectation, we have

$$\sum_{k=1}^{M} \frac{1}{2(L_g + \mu_k)} \mathbb{E}[\|\nabla f(x_k)\|_2^2] \le M\delta + \sum_{k=1}^{M} \frac{\sigma_h^2}{\mu_k} + \mathbb{E}[f(x_1) - f(x_{M+1})].$$

Since $\mu_k = O(L_g) \wedge (\mu_k = \Omega(L_g) \vee \sigma_h = 0)$ and $f(x_1) - f(x_{M+1}) \le f(x_1) - f_*$,

$$\sum_{k=1}^{M} \mathbb{E}[\|\nabla f(x_k)\|_2^2] \le O(L_g M\delta + M\sigma_h^2 + L_g(f(x_1) - f_*)).$$

Noting that

$$\mathbb{E}[\|\nabla f(x_R)\|_2^2 | \mathcal{F}_M] = \frac{1}{M} \sum_{k=1}^{M} \|\nabla f(x_k)\|_2^2,$$

we can conclude the proof of Theorem as follows,

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(x_R)\|_2^2] &= \mathbb{E}\left[\mathbb{E}[\|\nabla f(x_R)\|_2^2|\mathcal{F}_M]\right] = \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}[\|\nabla f(x_k)\|_2^2] \\
&\leq O\left(L_g\delta + \sigma_h^2 + \frac{L_g(f(x_1)-f_*)}{M}\right).
\end{aligned}
$$

□

Below is the proof of Proposition 2.

Proof of Proposition 2. It follows that

$$
\begin{aligned}
&\mathbb{E}[\|\nabla f(x_{k+1})\|_2^2|\mathcal{F}_k] \\
&\quad = \mathbb{E}[\|\nabla \phi_k(x_{k+1}) - (\nabla h(x_k) - v_h(x_k)) + \nabla h(x_k) - \nabla h(x_{k+1}) - \mu_k(x_{k+1}-x_k)\|_2^2|\mathcal{F}_k] \\
&\quad \leq 4\mathbb{E}[\|\nabla \phi_k(x_{k+1})\|_2^2 + \|\nabla h(x_k) - v_h(x_k)\|_2^2 + \|\nabla h(x_k) - \nabla h(x_{k+1})\|_2^2 + \mu_k^2\|x_{k+1}-x_k\|_2^2|\mathcal{F}_k] \\
&\quad \leq 4\sigma_h^2 + 4\mathbb{E}[\|\nabla \phi_k(x_{k+1})\|_2^2 + (\mu_k^2 + L_h^2)\|x_{k+1}-x_k\|_2^2|\mathcal{F}_k],
\end{aligned}
$$

where for the first inequality we used $\|\sum_{j=1}^{d}\alpha_j\|_2^2 \leq d\sum_{j=1}^{d}\|\alpha_j\|_2^2$ and for the second inequality we used Lipschitz smoothness of h. Since ϕ_k is $(L_g+\mu_k)$-smooth,

$$
\frac{1}{2(L_g+\mu_k)}\mathbb{E}[\|\nabla \phi_k(x_{k+1})\|_2^2|\mathcal{F}_k] \leq \mathbb{E}[\phi_k(x_{k+1}) - \phi_k^*|\mathcal{F}_k] \leq \delta.
$$

Thus, we conclude

$$
\mathbb{E}[\|\nabla f(x_{k+1})\|_2^2|\mathcal{F}_k] \leq 4\sigma_h^2 + 8(L_g+\mu_k)\delta + 4(\mu_k^2 + L_h^2)\mathbb{E}[\|x_{k+1}-x_k\|_2^2|\mathcal{F}_k].
$$

□

By combining Proposition 1 and 2, we prove Proposition 3.

Proof of Proposition 3. Noting that $\mu_k = O(L_h) \wedge \mu_k = \Omega(L_h)$, we have

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(x_{k+1})\|_2^2|\mathcal{F}_k] &\leq O\left((L_g+L_h)\delta + \sigma_h^2 + L_h^2\mathbb{E}[\|x_{k+1}-x_k\|_2^2|\mathcal{F}_k]\right) \\
&\leq O\left((L_g+L_h)\delta + \sigma_h^2 + L_h\mathbb{E}[f(x_k) - f(x_{k+1})|\mathcal{F}_k]\right),
\end{aligned}
$$

where for the first and second inequality we used Proposition 1 and 2, respectively. □

We give the proof of Theorem 2.

Proof of Theorem 2. Using Proposition 3 and $L_h = O(L_g)$, it follows that

$$
\mathbb{E}[\|\nabla f(x_{k+1})\|_2^2|\mathcal{F}_k] \leq O\left(L_g\delta + \sigma_h^2 + L_h\mathbb{E}[f(x_k) - f(x_{k+1})|\mathcal{F}_k]\right).
$$

This inequality resemble Proposition 1 up to the term $\mathbb{E}[\|x_{k+1}-x_k\|_2^2|\mathcal{F}_k]$, so that we can show the theorem in the same manner as Theorem 1. □

B The derivation of diagonal hessian approximation

To run AdaSPD with a diagonal hessian approximation for training BMs, we give $\mathrm{diag}(\nabla_\theta^2 h(\Theta))$, where h is the concave part of the log-likelihood of BMs. We only consider a parameter W_{ij} connecting a visible unit v^i and a hidden unit h^j because for the other parameters it can be shown in the same manner.

$$
\begin{aligned}
\nabla^2_{W_{ij}} h(\Theta) &= \nabla^2_{W_{ij}} \log \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) \\
&= \nabla_{W_{ij}} \left(\frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) v^i h^j}{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta))} \right) \\
&= \frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) (v^i h^j)^2}{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta))} - \left(\frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) v^i h^j}{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta))} \right)^2 \\
&= \frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) v^i h^j}{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta))} - \left(\frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) v^i h^j}{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta))} \right)^2 \\
&= \nabla_{W_{ij}} \log \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) - \left(\nabla_{W_{ij}} \log \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \Theta)) \right)^2 \\
&= \nabla_{W_{ij}} h(\Theta) - (\nabla_{W_{ij}} h(\Theta))^2,
\end{aligned}
$$

where we used $(v^i h^j)^2 = v^i h^j$ derived from the fact that v^i and h^j are binary units $\{0, 1\}$.

C Parameter settings for training RBMs and DBMs

In our experiments, we optimized L_2-penalized log-likelihoods of RBMs and DBMs. Here, we give all parameter settings used for AdaSPD. The damping parameter λ was fixed to 10^{-4}. The scales of metrics were set to $\mu = 10^{-4}$ for diagonal hessian approximation and $\mu \in \{10^{-5}, 10^{-3}, 10^{-1}\}$ for scalar matrix. The number of underlying solver iterations T and the suffix averaging parameter α were set as follows: $T = \lceil N/b \rceil, \alpha T = \lceil T/2 \rceil$, where N is the number of data points and b is a mini-batch size. The other parameters are listed in Table 1 for binarized MNIST dataset and Table 2 for CalTech101 Silhouettes.

Table 1: Parameter settings for binarized MNIST

Model	Minibatch-size b	PCD-k	Mean-field iter.	L_2-penalty	η
RBM-15	32	1	-	0	10^{-1}
RBM-25	32	3	-	0	10^{-1}
RBM-500	128	10	-	5×10^{-4}	10^{-1}
DBM-500-500-1000	128	10	10	3×10^{-4}	10^{-2}
DBM-500-500-500-1000	128	10	10	5×10^{-4}	10^{-2}

Table 2: Parameter settings for CalTech101 Silhouettes

Model	Minibatch-size b	PCD-k	Mean-field iter.	L_2-penalty	η
RBM-15	32	1	-	0	10^{-2}
RBM-25	32	3	-	0	10^{-2}
RBM-500	64	10	-	10^{-3}	10^{-2}