Appendix

A Proof of Theorem 1

Proof.

$$
\begin{aligned}
x_{t+1} =& \arg\min_x m_{1:t} \cdot x + \sum_{s=1}^{t} \frac{1}{2\alpha_s}(x - x_s)^T Q_s (x - x_s) \\
&+ \Psi_t(x) \\
=& \arg\min_x m_{1:t} \cdot x + \sum_{s=1}^{t} \frac{1}{2\alpha_s}\big(\|Q_s^{\frac{1}{2}} x\|_2^2 - 2x \cdot (Q_s x_s) \\
&+ \|Q_s^{\frac{1}{2}} x_s\|_2^2\big) + \Psi_t(x) \\
=& \arg\min_x (m_{1:t} - \sum_{s=1}^{t} \frac{Q_s}{\alpha_s} x_s) \cdot x + \sum_{s=1}^{t} \frac{1}{2\alpha_s}\|Q_s^{\frac{1}{2}} x\|_2^2 \\
&+ \Psi_t(x).
\end{aligned}
\tag{22}
$$

Define $z_{t-1} = m_{1:t-1} - \sum_{s=1}^{t-1} \frac{Q_s}{\alpha_s} x_s$ $(t \geq 2)$ and we can calculate z_t as

$$
z_t = z_{t-1} + m_t - \frac{Q_t}{\alpha_t} x_t, \quad t \geq 1. \tag{23}
$$

By substituting (23), (22) is simplified to be

$$
x_{t+1} = \arg\min_x z_t \cdot x + \sum_{s=1}^{t} \frac{Q_s}{2\alpha_s}\|x\|_2^2 + \Psi_t(x). \tag{24}
$$

By substituting $\Psi_t(x)$ (Eq. (5)) into (24), we get

$$
\begin{aligned}
x_{t+1} =& \arg\min_x z_t \cdot x + \sum_{g=1}^{G} \big(\lambda_1 \|x^g\|_1 + \lambda_{21}\sqrt{d_{x^g}} \\
& \|(\sum_{s=1}^{t} \frac{Q_s^g}{2\alpha_s} + \lambda_2 \mathbb{I})^{\frac{1}{2}} x^g\|_2\big) + \|(\sum_{s=1}^{t} \frac{Q_s}{2\alpha_s} + \lambda_2 \mathbb{I})^{\frac{1}{2}} x\|_2^2.
\end{aligned}
\tag{25}
$$

Since the objective of (25) is component-wise and element-wise, we can focus on the solution in one group, say g, and one entry, say i, in the g-th group. Let $\sum_{s=1}^{t} \frac{Q_s^g}{2\alpha_s} = \mathrm{diag}(\sigma_t^g)$ where $\sigma_t^g = (\sigma_{t,1}^g, \ldots, \sigma_{t,d_{x^g}}^g)$. The objective of (25) on $x_{t+1,i}^g$ is

$$
\Omega(x_{t+1,i}^g) = z_{t,i}^g x_{t+1,i}^g + \lambda_1 |x_{t+1,i}^g| + \Phi(x_{t+1,i}^g), \tag{26}
$$

where $\Phi(x_{t+1,i}^g) = \lambda_{21}\sqrt{d_{x^g}}\|(\sigma_{t,i}^g + \lambda_2)^{\frac{1}{2}} x_{t+1,i}^g\|_2 + \|(\sigma_{t,i}^g + \lambda_2)^{\frac{1}{2}} x_{t+1,i}^g\|_2^2$ is a non-negative function and $\Phi(x_{t+1,i}^g) = 0$ iff $x_{t+1,i}^g = 0$ for all $i \in \{1, \ldots, d_{x^g}\}$.
We discuss the optimal solution of (26) in three cases:

a) If $z^g_{t,i} = 0$, then $x^g_{t+1,i} = 0$.
b) If $z^g_{t,i} > 0$, then $x^g_{t+1,i} \le 0$. Otherwise, if $x^g_{t+1,i} > 0$, we have $\Omega(-x^g_{t+1,i}) < \Omega(x^g_{t+1,i})$, which contradicts the minimization value of $\Omega(x)$ on $x^g_{t+1,i}$.
Next, if $z^g_{t,i} \le \lambda_1$, then $x^g_{t+1,i} = 0$. Otherwise, if $x^g_{t+1,i} < 0$, we have $\Omega(x^g_{t+1,i}) = (z^g_{t,i} - \lambda_1)x^g_{t+1,i} + \Phi(x^{g,i}_{t+1}) > \Omega(0)$, which also contradicts the minimization value of $\Omega(x)$ on $x^g_{t+1,i}$.
Third, $z^g_{t,i} > \lambda_1$ ($\forall\ i = 1, \ldots, d_{x^g}$). The objective of (26) for the g-th group, $\Omega(x^g_{t+1})$, becomes
$$(z^g_t - \lambda_1 \mathbf{1}_{d_{x^g}}) \cdot x^g_{t+1} + \Phi(x^g_{t+1}).$$
c) If $z^g_{t,i} < 0$, the analysis is similar to b). We have $x^g_{t+1,i} \ge 0$. When $-z^g_{t,i} \le \lambda_1$, $x^g_{t+1,i} = 0$. When $-z^g_{t,i} > \lambda_1$ ($\forall\ i = 1, \ldots, d_{x^g}$), we have
$$\Omega(x^g_{t+1}) = (z^g_t + \lambda_1 \mathbf{1}_{d_{x^g}}) \cdot x^g_{t+1} + \Phi(x^g_{t+1}).$$

From a), b), c) above, we have
$$x^g_{t+1} = \arg\min_x -s^g_t \cdot x + \Phi(x), \tag{27}$$
where the i-th element of s^g_t is defined same as (9).
Define
$$y = (\mathrm{diag}(\sigma^g_t) + \lambda_2 \mathbb{I})^{\frac{1}{2}} x. \tag{28}$$
By substituting (28) into (27), we get
$$y^g_{t+1} = \arg\min_y -\tilde{s}^g_t \cdot y + \lambda_{21}\sqrt{d_{x^g}}\|y\|_2 + \|y\|^2_2, \tag{29}$$
where $\tilde{s}^g_t = (\mathrm{diag}(\sigma^g_t) + \lambda_2 \mathbb{I})^{-1} s^g_t$ which is defined same as (10). This is unconstrained non-smooth optimization problem. Its optimality condition (see [21], Section 27) states that y^g_{t+1} is an optimal solution if and only if there exists $\xi \in \partial\|y^g_{t+1}\|_2$ such that
$$-\tilde{s}^g_t + \lambda_{21}\sqrt{d_{x^g}}\xi + 2y^g_{t+1} = 0. \tag{30}$$
The subdifferential of $\|y\|_2$ is
$$\partial\|y\|_2 = \begin{cases} \{\zeta \in \mathbb{R}^{d_{x^g}} \mid -1 \le \zeta^{(i)} \le 1, i = 1, \ldots, d_{x^g}\} & \text{if } y = 0, \\ \frac{y}{\|y\|_2} & \text{if } y \ne 0. \end{cases}$$
Similarly to the analysis of ℓ_1-regularization, we discuss the solution of (30) in two different cases:

a) If $\|\tilde{s}^g_t\|_2 \le \lambda_{21}\sqrt{d_{x^g}}$, then $y^g_{t+1} = 0$ and $\xi = \frac{\tilde{s}^g_t}{\lambda_{21}\sqrt{d_{x^g}}} \in \partial\|0\|_2$ satisfy (30). We also show that there is no solution other than $y^g_{t+1} = 0$. Without loss of generality, we assume $y^g_{t+1,i} \ne 0$ for all $i \in \{1, \ldots, d_{x^g}\}$, then $\xi = \frac{y^g_{t+1}}{\|y^g_{t+1}\|_2}$, and
$$-\tilde{s}^g_t + \frac{\lambda_{21}\sqrt{d_{x^g}}}{\|y^g_{t+1}\|_2} y^g_{t+1} + 2y^g_{t+1} = 0. \tag{31}$$

From (31), we can derive

$$(\frac{\lambda_{21}\sqrt{d_{x^g}}}{\|y^g_{t+1}\|_2} + 2)\|y^g_{t+1}\|_2 = \|\tilde{s}^g_t\|_2.$$

Furthermore, we have

$$\|y^g_{t+1}\|_2 = \frac{1}{2}(\|\tilde{s}^g_t\|_2 - \lambda_{21}\sqrt{d_{x^g}}), \tag{32}$$

where $\|y^g_{t+1}\|_2 > 0$ and $\|\tilde{s}^g_t\|_2 - \lambda_{21}\sqrt{d_{x^g}} \le 0$ contradict each other.

b) If $\|\tilde{s}^g_t\|_2 > \lambda_{21}\sqrt{d_{x^g}}$, then from (31) and (32), we get

$$y^g_{t+1} = \frac{1}{2}(1 - \frac{\lambda_{21}\sqrt{d_{x^g}}}{\|\tilde{s}^g_t\|_2})\tilde{s}^g_t. \tag{33}$$

We replace y^g_{t+1} of (33) by x^g_{t+1} using (28), then we have

$$\begin{aligned} x^g_{t+1} &= (\mathrm{diag}(\sigma^g_t) + \lambda_2\mathbb{I})^{-\frac{1}{2}} y^g_{t+1} \\ &= (2\mathrm{diag}(\sigma^g_t) + 2\lambda_2\mathbb{I})^{-1}(1 - \frac{\lambda_{21}\sqrt{d_{x^g}}}{\|\tilde{s}^g_t\|_2})s^g_t \\ &= (\sum_{s=1}^{t}\frac{Q_s}{\alpha_s} + 2\lambda_2\mathbb{I})^{-1}(1 - \frac{\lambda_{21}\sqrt{d_{x^g}}}{\|\tilde{s}^g_t\|_2})s^g_t. \end{aligned} \tag{34}$$

Combine a) and b) above, we finish the proof.

B Proof of Theorem 2

Proof. We use the method of induction.

a) When $t = 1$, then Algorithm 1 becomes

$$\begin{aligned} Q_1 &= \alpha_1(\frac{\sqrt{V_1}}{\alpha_1} - \frac{\sqrt{V_0}}{\alpha_0}) = \sqrt{V_1}, \\ z_1 &= z_0 + m_1 - \frac{Q_1}{\alpha_1}x_1 = m_1 - \frac{\sqrt{V_1}}{\alpha_1}x_1, \\ s_1 &= -z_1 = \frac{\sqrt{V_1}}{\alpha_1}x_1 - m_1, \\ x_2 &= (\frac{\sqrt{V_1}}{\alpha_1})^{-1}s_1 = x_1 - \alpha_1\frac{m_1}{\sqrt{V_1}}, \end{aligned}$$

which equals to Eq. (1).

b) Assume $t = T$, Eq. (35) are true.

$$\begin{aligned} z_T &= m_T - \frac{\sqrt{V_T}}{\alpha_T}x_T, \\ x_{T+1} &= x_T - \alpha_T\frac{m_T}{\sqrt{V_T}}. \end{aligned} \tag{35}$$

For $t = T+1$, we have

$$\begin{aligned}
z_{T+1} &= z_T + m_{T+1} - \frac{Q_{T+1}}{\alpha_{T+1}} x_{T+1} \\
&= m_T - \frac{\sqrt{V_T}}{\alpha_T} x_T + m_{T+1} - \frac{Q_{T+1}}{\alpha_{T+1}} x_{T+1} \\
&= m_T - \frac{\sqrt{V_T}}{\alpha_T}(x_{T+1} + \alpha_T \frac{m_T}{\sqrt{V_T}}) + m_{T+1} - \frac{Q_{T+1}}{\alpha_{T+1}} x_{T+1} \\
&= m_{T+1} - (\frac{\sqrt{V_T}}{\alpha_T} + \frac{Q_{T+1}}{\alpha_{T+1}}) x_{T+1} \\
&= m_{T+1} - \frac{\sqrt{V_{T+1}}}{\alpha_{T+1}} x_{T+1}, \\
x_{T+2} &= (\frac{\sqrt{V_{T+1}}}{\alpha_{T+1}})^{-1} s_{T+1} = -(\frac{\sqrt{V_{T+1}}}{\alpha_{T+1}})^{-1} z_{T+1} \\
&= x_{T+1} - \alpha_T \frac{m_{T+1}}{\sqrt{V_{T+1}}}.
\end{aligned}$$

Hence, we complete the proof.

C Proof of Theorem 3

Proof. Let

$$h_t(x) = \begin{cases} \sum_{s=1}^{t} \frac{1}{2\alpha_s} \|Q_s^{\frac{1}{2}}(x - x_s)\|_2^2 \ \forall\ t \in [T], \\ \frac{1}{2}\|x - c\|_2^2 \qquad\qquad\qquad t = 0. \end{cases}$$

It is easy to verify that for all $t \in [T]$, $h_t(x)$ is 1-strongly convex with respect to $\|\cdot\|_{\sqrt{V_t}/\alpha_t}$ which $\frac{\sqrt{V_t}}{\alpha_t} = \sum_{s=1}^{t} \frac{Q_s}{\alpha_s}$, and $h_0(x)$ is 1-strongly convex with respect to $\|\cdot\|_2$.
From (7), we have

$$\begin{aligned}
\mathcal{R}_T &= \sum_{t=1}^{T} (f_t(x_t) - f_t(x^*)) \le \sum_{t=1}^{T} \langle g_t, x_t - x^* \rangle \\
&= \sum_{t=1}^{T} \langle m_t - \gamma m_{t-1}, x_t - x^* \rangle \le \sum_{t=1}^{T} \langle m_t, x_t - x^* \rangle \\
&= \sum_{t=1}^{T} \langle m_t, x_t \rangle + \Psi_T(x^*) + h_T(x^*) + \big(\sum_{t=1}^{T} \langle -m_t, x^* \rangle \\
&\quad - \Psi_T(x^*) - h_T(x^*) \big) \\
&\le \sum_{t=1}^{T} \langle m_t, x_t \rangle + \Psi_T(x^*) + h_T(x^*) + \sup_{x \in \mathcal{Q}} \big\{ \langle -m_{1:T}, x \rangle \\
&\quad - \Psi_T(x) - h_T(x) \big\},
\end{aligned} \tag{36}$$

where in the first and second inequality above, we use the convexity of $f_t(x)$ and the condition (12) respectively.
We define $h_t^*(u)$ to be the conjugate dual of $\Psi_t(x) + h_t(x)$:

$$h_t^*(u) = \sup_{x \in \mathcal{Q}} \left\{ \langle u, x \rangle - \Psi_t(x) - h_t(x) \right\}, \quad t \geq 0,$$

where $\Psi_0(x) = 0$. Since $h_t(x)$ is 1-strongly convex with respect to the norm $\|\cdot\|_{h_t}$, the function h_t^* has 1-Lipschitz continuous gradients with respect to $\|\cdot\|_{h_t^*}$ (see, [14], Theorem 1):

$$\|\nabla h_t^*(u_1) - \nabla h_t^*(u_2)\|_{h_t} \leq \|u_1 - u_2\|_{h_t^*}, \tag{37}$$

and

$$\nabla h_t^*(u) = \arg\min_{x \in \mathcal{Q}} \left\{ -\langle u, x \rangle + \Psi_t(x) + h_t(x) \right\}. \tag{38}$$

As a trivial corollary of (37), we have the following inequality:

$$h_t^*(u + \delta) \leq h_t^*(u) + \langle \nabla h_t^*(u), \delta \rangle + \frac{1}{2} \|\delta\|_{h_t^*}^2. \tag{39}$$

Since $h_{t+1}(x) \geq h_t(x)$ and $\Psi_{t+1}(x) \geq \Psi_t(x)$, from (38), (39), (6), we have

$$\begin{aligned} h_T^*(-m_{1:T}) \leq & h_{T-1}^*(-m_{1:T}) \\ \leq & h_{T-1}^*(-m_{1:T-1}) - \left\langle \nabla h_{T-1}^*(-m_{1:T-1}), m_T \right\rangle \\ & + \frac{1}{2} \|m_T\|_{h_{T-1}^*}^2 \\ \leq & h_{T-2}^*(-m_{1:T-1}) - \langle x_T, m_T \rangle + \frac{1}{2} \|m_T\|_{h_{T-1}^*}^2 \\ \leq & h_0^*(0) - \langle \nabla h_0^*(0), m_1 \rangle - \sum_{t=2}^{T} \langle x_t, m_t \rangle \\ & + \frac{1}{2} \sum_{t=2}^{T} \|m_t\|_{h_{t-1}^*}^2 \\ = & -\sum_{t=1}^{T} \langle x_t, m_t \rangle + \frac{1}{2} \sum_{t=1}^{T} \|m_t\|_{h_{t-1}^*}^2 . \end{aligned} \tag{40}$$

where the last equality above follows from $h_0^*(0) = 0$ and (11) which deduces $x_1 = \nabla h_0^*(0)$.
By substituting (40), (36) becomes

$$\begin{aligned} \mathcal{R}_T \leq & \sum_{t=1}^{T} \langle m_t, x_t \rangle + \Psi_T(x^*) + h_T(x^*) + h_T^*(-m_{1:T}) \\ \leq & \Psi_T(x^*) + h_T(x^*) + \frac{1}{2} \sum_{t=1}^{T} \|m_t\|_{h_{t-1}^*}^2 . \end{aligned} \tag{41}$$

D Additional Proofs

D.1 Proof of Lemma 1

Proof. Let $V_t = \mathrm{diag}(\sigma_t)$ where σ_t is the vector of the diagonal elements of V_t. For i-th entry of σ_t, by substituting (13) into (15), we have

$$\begin{aligned}
\sigma_{t,i} &= g_{t,i}^2 + \eta\sigma_{t-1,i} \\
&= (m_{t,i} - \gamma m_{t-1,i})^2 + \eta g_{t-1,i}^2 + \eta^2\sigma_{t-2,i} \\
&= \sum_{s=1}^{t} \eta^{t-s}(m_{s,i} - \gamma m_{s-1,i})^2 \\
&\geq \sum_{s=1}^{t} \eta^{t-s}(1-\gamma)(m_{s,i}^2 - \gamma m_{s-1,i}^2) \\
&= (1-\gamma)\big(m_{t,i}^2 + (\eta-\gamma)\sum_{s=1}^{t-1}\eta^{t-s-1}m_{s,i}^2\big).
\end{aligned} \tag{42}$$

Next, we will discuss the value of η in two cases.

a) $\eta = 1$. From (42), we have

$$\begin{aligned}
\sigma_{t,i} &\geq (1-\gamma)\big(m_{t,i}^2 + (1-\gamma)\sum_{s=1}^{t-1} m_{s,i}^2\big) \\
&> (1-\gamma)^2 \sum_{s=1}^{t} m_{s,i}^2 \\
&\geq (1-\nu)^2 \sum_{s=1}^{t} m_{s,i}^2.
\end{aligned} \tag{43}$$

Recalling the definition of $M_{t,i}$ in Section 1.5, from (43), we have

$$\sum_{t=1}^{T} \frac{m_{t,i}^2}{\sqrt{\sigma_{t,i}}} < \frac{1}{1-\nu}\sum_{t=1}^{T}\frac{m_{t,i}^2}{\|M_{t,i}\|_2} \leq \frac{2}{1-\nu}\|M_{T,i}\|_2,$$

where the last inequality above follows from Appendix C of [4]. Therefore, we get

$$\begin{aligned}
\sum_{t=1}^{T} \|m_t\|^2_{(\frac{\sqrt{V_t}}{\alpha_t})^{-1}} &= \alpha\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{m_{t,i}^2}{\sqrt{\sigma_{t,i}}} \\
&< \frac{2\alpha}{1-\nu}\sum_{i=1}^{d}\|M_{T,i}\|_2.
\end{aligned} \tag{44}$$

b) $\eta < 1$. We assume $\eta \geq \gamma$ and $\kappa V_t \succeq V_{t-1}$ where $\kappa < 1$, then we have

$$\sum_{s=1}^{t} \kappa^{t-s}\sigma_{t,i} \geq \sum_{s=1}^{t}\sigma_{s,i} \geq (1-\gamma)\sum_{s=1}^{t} m_{s,i}^2.$$

Hence, we get

$$\begin{aligned}
\sigma_{t,i} &\geq \frac{1-\kappa}{1-\kappa^t}(1-\gamma)\sum_{s=1}^{t} m_{s,i}^2 \\
&> (1-\kappa)(1-\gamma)\sum_{s=1}^{t} m_{s,i}^2 \\
&\geq (1-\nu)^2\sum_{s=1}^{t} m_{s,i}^2,
\end{aligned} \tag{45}$$

which deduces the same conclusion (44) of a).

Combine a) and b), we complete the proof.

D.2 Proof of Corollary 1

Proof. From the definition of m_t (13), V_t (15), we have

$$\begin{aligned}
|m_{t,i}| &= |\sum_{s=1}^{t} \gamma^{t-s} g_{s,i}| \leq \frac{1-\gamma^t}{1-\gamma} G < \frac{G}{1-\gamma} \leq \frac{G}{1-\nu}, \\
|\sigma_{t,i}| &= |\sum_{s=1}^{t} \eta^{t-s} g_{s,i}^2| \leq tG^2.
\end{aligned}$$

Hence, we have

$$\Psi_T(x^*) \leq \lambda_1 dD_1 + \lambda_{21} dD_1 (\frac{\sqrt{T}G}{2\alpha} + \lambda_2)^{\frac{1}{2}} + \lambda_2 dD_1^2, \tag{46}$$

$$h_T(x^*) \leq \frac{dD_2^2 G}{2\alpha}\sqrt{T}, \tag{47}$$

$$\frac{1}{2}\sum_{t=1}^{T} \|m_t\|_{h_{t-1}^*}^2 < \frac{\alpha}{1-\nu}\sum_{i=1}^{d} \frac{\sqrt{T}G}{1-\nu} = \frac{d\alpha G}{(1-\nu)^2}\sqrt{T}. \tag{48}$$

Combining (46), (47), (48), we complete the proof.

E Additional Experimental Results

Table 5. The learning rates of the optimizers of three datasets.

Optimizer	MLP	OPNN	DCN
Adam/Group Adam	1e-4	1e-4	1e-3
AdaGrad/Group AdaGrad	1e-2	1e-2	1e-2

Table 6. The regularization terms of GROUP ADAM of three datasets.

Dataset	λ_1	λ_{21}	λ_2
MLP	5e-3	1e-2	1e-5
OPNN	8e-5	1e-5	1e-5
DCN	4e-4	5e-4	1e-5

Table 7. The regularization terms of GROUP ADAGRAD of three datasets.

Dataset	λ_1	λ_{21}	λ_2
MLP	0	1e-2	1e-5
OPNN	8e-5	8e-5	1e-5
DCN	0	4e-3	1e-5

Table 8. Fine-tuned schedule of magnitude pruning. The best AUC for each dataset on each sparsity level is bolded.

Ratio of Samples	MLP		OPNN		DCN	
	Sparsity	AUC	Sparsity	AUC	Sparsity	AUC
30%	0.974	0.7454	0.078	0.6809	0.518	0.8016
20%		**0.7457**		0.7259		0.8015
10%		0.7439		0.7160		0.8015
0%		0.7452		**0.7551**		**0.8019**
30%	0.864	0.7454	0.039	0.6356	0.062	0.8005
20%		0.7441		0.6383		0.7977
10%		0.7453		0.6678		0.7884
0%		**0.7464**		**0.7491**		**0.8018**
30%	0.701	0.7443	0.032	0.6826	0.018	0.7883
20%		0.7452		0.6618		0.7969
10%		0.7449		0.6604		0.7800
0%		**0.7452**		**0.7465**		**0.8017**
30%	0.132	**0.7457**	0.018	0.6318	0.0042	0.7400
20%		0.7428		0.6419		0.7236
10%		0.7450		0.6207		0.6864
0%		0.7452		**0.7509**		**0.7995**
30%	0.038	**0.7444**	0.0092	0.5934	0.0025	0.7442
20%		0.7442		0.6003		0.7105
10%		0.7437		0.6355		0.6925
0%		0.7430		**0.7396**		**0.7972**

Table 9. The ℓ_{21}-regularization terms of the optimizer using s_t and $\tilde{s}_t$.

Method	λ_{21}				
s_t	0	1e-4	2.5e-4	5e-4	1e-3
	2.5e-3	5e-3	7.5e-3	1e-2	2.5e-2
$\tilde{s}_t$	0	0.05	0.075	0.1	0.125
	0.15	0.175	0.2	0.225	0.25