A. Continuous analysis

To motivate our proofs for the theorems in main text, let us first elaborate the continuous cases. Then we will extend our analysis to the discrete circumstances. One can safely skip this part and go directly to Section C for the missing proofs in main text, which is self-consistent.

Continuous optimization paths To ease notations and preliminaries, in this part we only discuss gradient descent (GD) and Nesterov's accelerated gradient descent (NGD), and their strong continuous approximation via ordinary differential equations (ODEs). For SGD and NSGD, existing works show that there are weak continuous approximation by stochastic differential equations (SDEs) (Hu et al., 2017a;b; Li et al., 2017). Our analysis can be extended to SDEs, but we believe it serves better to motivate our discrete proofs by focusing on ODEs.

We consider loss $L(w)$ and ℓ_2-regularizer $R(w) = \frac{1}{2}\left\|w\right\|_2^2$. Let the learning rate $\eta \to 0$, the path of $L(w)$ optimized by GD converges to the following ODE (Yang et al., 2018)

$$\mathrm{d}w_t = -\nabla L(w_t)\mathrm{d}t.$$

Similarly the continuous GD optimization path of regularized loss admits

$$\mathrm{d}\hat{w}_t = -\left(\nabla L(\hat{w}_t) + \lambda \hat{w}_t\right)\mathrm{d}t.$$

As for NGD, Su et al. (2014); Yang et al. (2018) show if the loss is α-strongly convex, then the NGD optimization path converges to

$$w_t'' + 2\sqrt{\alpha}w_t' + L'(w_t) = 0.$$

Since $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\left\|\hat{w}\right\|_2^2$ is $(\alpha+\lambda)$-strongly convex, the NGD path of the regularized loss satisfies

$$\hat{w}_t'' + 2\sqrt{\alpha+\lambda}\hat{w}_t' + L'(\hat{w}_t) + \lambda\hat{w}_t = 0.$$

Continuous weighting scheme We define the continuous weighting scheme as

$$p_t \ge 0, \quad t \ge 0, \quad P_t = \int_0^t p(s)\mathrm{d}s, \quad \lim_{t\to\infty} P_t = 1.$$

Lemma 1. *Given two continuous dynamic $x_t,\ \hat{x}_t,\ t \ge 0$. Let $\tilde{x}_t = P_t^{-1}\int_0^t p_s x_s \mathrm{d}s$. Suppose $x_0 = \hat{x}_0 = 0$. If the continuous weighting scheme P_t satisfies*

$$\mathrm{d}\hat{x}_t = (1-P_t)\mathrm{d}x_t, \quad t \ge 0,$$

then we have

$$P_t(x_t - \tilde{x}_t) = x_t - \hat{x}_t, \quad t \ge 0,$$

and

$$\hat{x}_t - \tilde{x}_t = (1-P_t)(x_t - \tilde{x}_t), \quad t \ge 0.$$

Proof. By definition we have for $t \ge 0$,

$$\begin{aligned}\tilde{x}_t = P_t^{-1}\int_0^t p_s x_s \mathrm{d}s = P_t^{-1}\left(x_s P_s|_0^t - \int_0^t P_s \mathrm{d}x_s\right) = x_t - P_t^{-1}\int_0^t P_s\mathrm{d}x_s \\ = x_t - P_t^{-1}\left(x_t - \int_0^t (1-P_s)\mathrm{d}x_s\right) = x_t - P_t^{-1}\left(x_t - \int_0^t \mathrm{d}\hat{x}_s\right) \\ = x_t - P_t^{-1}\left(x_t - \hat{x}_t\right).\end{aligned}$$

Thus

$$P_t(x_t - \tilde{x}_t) = x_t - \hat{x}_t,$$

and

$$\hat{x}_t - \tilde{x}_t = x_t - P_t(x_t - \tilde{x}_t) - \tilde{x}_t = (1-P_t)(x_t - \tilde{x}_t).$$

∎

A.1. Continuous Theorem 1

Consider linear regression problem $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^\top x_i - y_i\right\|_2^2 = \frac{1}{2}w^\top \Sigma w - w^\top a + \text{const}$, and ℓ_2-regularizer $R(w) = \frac{1}{2}\left\|w\right\|_2^2$. Assume the initial condition $w_0 = \hat{w}_0 = 0$, then the GD dynamics for the unregularized and regularized losses are

$$\begin{aligned} \mathrm{d}w_t &= -\left(\Sigma w_t - a\right)\mathrm{d}t, \quad w_0 = 0, \\ \mathrm{d}\hat{w}_t &= -\left(\Sigma \hat{w}_t - a + \lambda \hat{w}_t\right)\mathrm{d}t, \quad \hat{w}_0 = 0. \end{aligned}$$

The ODEs are solved by

$$w_t = \left(I - e^{-\Sigma t}\right)\Sigma^{-1}a, \quad \hat{w}_t = \left(I - e^{-(\Sigma+\lambda I)t}\right)(\Sigma + \lambda I)^{-1}a.$$

Now let the continuous weighting scheme be

$$P_t = 1 - e^{\lambda t},$$

then we have

$$\mathrm{d}\hat{w}_t = (1 - P_t)\mathrm{d}w_t,$$

thus by Lemma 1 we obtain

$$\hat{w}_t - \tilde{w}_t = (1 - P_t)(w_t - \tilde{w}_t),$$

which proves the continuous version of Theorem 1.

A.2. Continuous Theorem 3

Consider linear regression problem $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^\top x_i - y_i\right\|_2^2 = \frac{1}{2}w^\top \Sigma w - w^\top a + \text{const}$, and ℓ_2-regularizer $R(w) = \frac{1}{2}\left\|w\right\|_2^2$. Assume the initial condition $w_0 = w_0' = 0$ and $\hat{w}_0 = \hat{w}_0' = 0$. Then the unregularized and regularized NGD dynamics are

$$w_t'' + 2\sqrt{\alpha}w_t' + \Sigma w_t - a = 0, \quad w_0 = w_0' = 0, \tag{7}$$

$$\hat{w}_t'' + 2\sqrt{\alpha+\lambda}\hat{w}_t' + (\Sigma + \lambda)\hat{w}_t - a = 0, \quad \hat{w}_0 = \hat{w}_0' = 0. \tag{8}$$

We first solve the order-2 ODE Eq. (7) in the canonical way, and then obtain the solution of Eq. (8) similarly. To do so, let's firstly ignore the constant term and solve the homogenous ODE of Eq. (7), and obtain two general solutions of the homogenous equation as

$$w_{t,1} = e^{\sqrt{\alpha}t}\cos\sqrt{\Sigma - \alpha}t, \quad w_{t,2} = e^{\sqrt{\alpha}t}\sin\sqrt{\Sigma - \alpha}t.$$

Then we guess a particular solution of Eq. (7) as $w_{t,0} = \Sigma^{-1}a$. Thus the general solution of ODE (7) can be decomposed as $w_t = \lambda_1 w_{t,1} + \lambda_2 w_{t,2} + w_{t,0}$. Consider the initial conditions $w_0 = w_0' = 0$, we obtain $\lambda_1 = -\Sigma^{-1}a$, $\lambda_2 = -\Sigma^{-1}a\sqrt{(\Sigma - \alpha)^{-1}\alpha}$. Thus the solution of Eq. (7) is

$$\begin{aligned} w_t &= \Sigma^{-1}a\left(1 - e^{-\sqrt{\alpha}t}\cos\sqrt{\Sigma - \alpha}t - \sqrt{\alpha(\Sigma - \alpha)^{-1}}e^{-\sqrt{\alpha}t}\sin\sqrt{\Sigma - \alpha}t\right), \\ w_t' &= a\sqrt{(\Sigma - \alpha)^{-1}}e^{-\sqrt{\alpha}t}\sin\sqrt{\Sigma - \alpha}t. \end{aligned} \tag{9}$$

Repeat these procedures, Eq. (9) is solved by

$$\begin{aligned} \hat{w}_t &= (\Sigma + \lambda)^{-1}a\left(1 - e^{-\sqrt{\alpha+\lambda}t}\cos\sqrt{\Sigma - \alpha}t - \sqrt{(\alpha + \lambda)(\Sigma - \alpha)^{-1}}e^{-\sqrt{\alpha+\lambda}t}\sin\sqrt{\Sigma - \alpha}t\right), \\ \hat{w}_t' &= a\sqrt{(\Sigma - \alpha)^{-1}}e^{-\sqrt{\alpha+\lambda}t}\sin\sqrt{\Sigma - \alpha}t. \end{aligned} \tag{10}$$

Now let the continuous weighting scheme be

$$P_t = 1 - e^{-(\sqrt{\alpha+\lambda} - \sqrt{\lambda})t},$$

then we have

$$\mathrm{d}\hat{w}_t = (1 - P_t)\mathrm{d}w_t,$$

thus by Lemma 1 we obtain

$$\hat{w}_t - \tilde{w}_t = (1 - P_t)(w_t - \tilde{w}_t),$$

which proves the continuous version of Theorem 3.

A.3. Continuous Theorem 4

Consider an α-strongly convex and β-smooth loss function $L(w)$, and ℓ_2-regularizer. Without loss of generality assume the minimum of $L(w)$ satisfies $w_* > w_0 = 0$. Then by Lemma 3 we have

$$\alpha w - b \le \nabla L(w) \le \beta w - b, \quad \forall w \in (0, w_*),$$

where $b = -\nabla L(0)$, and "$\le$" is defined entry-wisely. We study the continuous optimization paths caused by GD.

Consider the following three dynamics:

$$\mathrm{d}w_t = -\nabla L(w_t)\mathrm{d}t, \quad \mathrm{d}u_t = -(\alpha u_t - b)\mathrm{d}t, \quad \mathrm{d}v_t = -(\beta v_t - b)\mathrm{d}t, \quad w_0 = u_0 = v_0 = 0.$$

By the comparison theorem of ODEs (Gronwall's inequality), and solution of linear ODEs, we claim that for all $t > 0$,

$$v_t \le w_t \le u_t, \quad u_t = \frac{b}{\alpha}(1 - e^{-\alpha t}), \quad v_t = \frac{b}{\beta}(1 - e^{-\beta t}). \tag{11}$$

In a similar manner, for the following three dynamics of regularized loss:

$$\mathrm{d}\hat{w}_{t,\lambda} = -(\nabla L(\hat{w}_{t,\lambda}) + \lambda \hat{w}_{t,\lambda})\mathrm{d}t, \qquad \mathrm{d}\hat{u}_{t,\lambda} = -((\lambda+\alpha)\hat{u}_{t,\lambda} - b)\mathrm{d}t, \quad \mathrm{d}\hat{v}_{t,\lambda} = -((\lambda+\beta)\hat{v}_{t,\lambda} - b)\mathrm{d}t,$$

where $\hat{w}_{0,\lambda} = \hat{u}_{0,\lambda} = \hat{v}_{0,\lambda} = 0$. Similarly we have for all $t > 0$,

$$\hat{v}_{t,\lambda} \le \hat{w}_{t,\lambda} \le \hat{u}_{t,\lambda}, \quad \hat{u}_{t,\lambda} = \frac{b}{\lambda+\alpha}(1 - e^{-(\lambda+\alpha)t}), \quad \hat{v}_{t,\lambda} = \frac{b}{\lambda+\beta}(1 - e^{-(\lambda+\beta)t}).$$

For the continuous weighting scheme

$$P_t = 1 - e^{-\zeta t}, \quad p_t = \zeta e^{-\zeta t}, \quad t \ge 0, \quad \zeta > 0,$$

the averaged solution is defined as $\tilde{w}_t = P_t^{-1}\int_0^t p_t w_t \mathrm{d}t = w_t - P_t^{-1}\int_0^t P_s \mathrm{d}w_s$, similar there are $\tilde{u}_t, \tilde{v}_t$. Thanks to Eq. (11) and p_t being non-negative, we have $\tilde{v}_t \le \tilde{w}_t \le \tilde{u}_t$. Let

$$\lambda_1 = \zeta + \beta - \alpha, \quad \lambda_2 = \zeta + \alpha - \beta,$$

then

$$\begin{aligned} P_t(u_t - \tilde{u}_t) &= \int_0^t P_s \mathrm{d}u_s = \int_0^t (1 - e^{-(\lambda_2+\beta-\alpha)s}) b e^{-\alpha s}\mathrm{d}t = b\int_0^t e^{-\alpha s} - e^{-(\beta+\lambda_2)s}\mathrm{d}s \\ &= b\left(\frac{1}{\alpha}(1 - e^{-\alpha t}) - \frac{1}{\lambda_2+\beta}(1 - e^{-(\lambda_2+\beta)t})\right) = u_t - \hat{v}_{t,\lambda_2}. \end{aligned}$$

Thus

$$\tilde{w}_t - \hat{w}_{t,\lambda_2} \le \tilde{u}_t - \hat{v}_{t,\lambda_2} = \tilde{u}_t - u_t + P_t(u_t - \tilde{u}_t) = (1 - P_t)(\tilde{u}_t - u_t).$$

Similarly, since

$$\begin{aligned} P_t(v_t - \tilde{v}_t) &= \int_0^t P_s \mathrm{d}v_s = \int_0^t (1 - e^{-(\lambda_1-\beta+\alpha)s}) b e^{-\beta s}\mathrm{d}t = b\int_0^t e^{-\beta s} - e^{-(\alpha+\lambda_1)s}\mathrm{d}s \\ &= b\left(\frac{1}{\beta}(1 - e^{-\beta t}) - \frac{1}{\lambda_1+\alpha}(1 - e^{-(\lambda_1+\alpha)t})\right) = v_t - \hat{u}_{t,\lambda_1}, \end{aligned}$$

we can obtain a lower bound as

$$\tilde{w}_t - \hat{w}_{t,\lambda_1} \ge \tilde{v}_t - \hat{u}_{t,\lambda_1} = \tilde{v}_t - v_t + P_t(v_t - \tilde{v}_t) = (1 - P_t)(\tilde{v}_t - v_t).$$

These inequalities give us

$$\hat{w}_{t,\lambda_1} + (1 - P_t)(\tilde{v}_t - v_t) \le \tilde{w}_t \le \hat{w}_{t,\lambda_2} + (1 - P_t)(\tilde{u}_t - u_t),$$

which proves the continuous version of Theorem 4.

B. Technical Lemmas

Lemma 2. *Consider two series $\{x_k\}_{k=0}^{\infty}$, $\{\hat{x}_k\}_{k=0}^{\infty}$, and a weighting scheme $\{p_k\}_{k=0}^{\infty}$ such that $\sum_{k=0}^{\infty} p_k = 1,\ p_k \ge 0$, $P_k = \sum_{i=1}^{k} p_i$. Let $\tilde{x}_k := P_k^{-1} \sum_{i=0}^{k} p_i x_i$. Suppose $x_0 = \hat{x}_0 = 0$. Suppose the weighting scheme P_k satisfies*

$$\hat{x}_{k+1} - \hat{x}_k = (1 - P_k)(x_{k+1} - x_k), \quad k \ge 0.$$

Then we have

$$P_k(x_k - \tilde{x}_k) = x_k - \hat{x}_k, \quad k \ge 0,$$

and

$$\hat{x}_k - \tilde{x}_k = (1 - P_k)\,(x_k - \tilde{x}_k), \quad k \ge 0.$$

More generally, the weighting scheme $\{p_k\}_{k=0}^{\infty}$ could be a series of positive semi-definite matrix where

$$\lim_{k \to +\infty} P_k = I, \quad 0 \preceq P_k \preceq I, \quad p_k = P_k - P_{k-1}.$$

Proof. By definition we know $p_0 = P_0,\ p_k = P_k - P_{k-1},\ k \ge 1$, and

$$\begin{aligned} P_k \tilde{x}_k =& \sum_{i=1}^{k} p_i x_i = \sum_{i=1}^{k} (P_i - P_{i-1}) x_i = \sum_{i=1}^{k} P_i x_i - \sum_{i=1}^{k} P_{i-1} x_i \\ =& P_k x_k + \sum_{i=1}^{k} P_{i-1} x_{i-1} - \sum_{i=1}^{k} P_{i-1} x_i = P_k x_k - \sum_{i=1}^{k} P_{i-1}(x_i - x_{i-1}). \end{aligned}$$

Therefore

$$\begin{aligned} P_k(x_k - \tilde{x}_k) =& \sum_{i=1}^{k} P_{i-1}(x_i - x_{i-1}) = \sum_{i=1}^{k} (x_i - x_{i-1}) - \sum_{i=1}^{k} (1 - P_{i-1})(x_i - x_{i-1}) \\ =& x_k - \sum_{i=1}^{k} (1 - P_{i-1})(x_i - x_{i-1}). \end{aligned}$$

Now use the assumption, we obtain

$$P_k(x_k - \tilde{x}_k) = x_k - \sum_{i=1}^{k} (\hat{x}_i - \hat{x}_{i-1}) = x_k - \hat{x}_k, \quad k \ge 1.$$

Thus we have

$$\hat{x}_k - \tilde{x}_k = x_k - P_k(x_k - \tilde{x}_k) - \tilde{x}_k = (1 - P_k)\,(x_k - \tilde{x}_k), \quad k \ge 1.$$

One can directly verify that the above equation also holds for $k = 0$, which concludes our proof. $\square$

Lemma 3. *Let $x \in \mathbb{R}$. Let $f(x)$ be α-strongly convex and β-smooth, $0 < \alpha \le \beta$. Let $f(x)$ be lower bounded, then $x_* = \arg\min_{x \in \mathbb{R}} f(x)$ exists. Consider GD with learning rate $\eta \in (0, \frac{1}{\beta})$, the optimization path $\{x_k\}_{k=0}^{+\infty}$ is given by*

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

If $x_0 < x_*$, then we have

1. *For all $k > 0$, $x_k \in (x_0, x_*)$.*
2. *For all $x \in (x_0, x_*)$, we have $\beta(x - x_*) \le \nabla f(x) \le \alpha(x - x_*)$.*
3. *For all $x \in (x_0, x_*)$, we have $\alpha(x - x_0) + \nabla f(x_0) \le \nabla f(x) \le \beta(x - x_0) + \nabla f(x_0)$.*

Similarly if $x_0 > x_*$, then we have

1. *For all $k > 0$, $x_k \in (x_*, x_0)$.*

2. *For all $x \in (x_*, x_0)$, we have $\alpha(x - x_*) \leq \nabla f(x) \leq \beta(x - x_*)$.*

3. *For all $x \in (x_*, x_0)$, we have $\beta(x - x_0) + \nabla f(x_0) \leq \nabla f(x) \leq \alpha(x - x_0) + \nabla f(x_0)$.*

Proof. We only prove Lemma 3 in case of $x_0 < x_*$. The other case is true in a similar manner.

To prove the first conclusion we only need to show that $x_0 < x_1 < x_*$, then recursively we obtain $x_0 < x_1 < \cdots < x_k < x_*$. Note that $\nabla f(x_*) = 0$. Since $f(x)$ is α-strongly convex and β-smooth, we have (Zhou, 2018)

$$\alpha(x-y)^2 \leq (\nabla f(x) - \nabla f(y))\,(x-y) \leq \beta(x-y)^2.$$

Thus $\alpha(x_* - x_0)^2 \leq -\nabla f(x_0)(x_* - x_0) \leq \beta(x_* - x_0)^2$. Now by the assumption that $x_0 < x_*$, we obtain $0 < \alpha(x_* - x_0) \leq -\nabla f(x_0) \leq \beta(x_* - x_0)$. Hence

$$\begin{aligned}
&x_1 = x_0 - \eta\nabla f(x_0) > x_0 \\
&x_1 = x_0 - \eta\nabla f(x_0) < x_0 + \eta\beta(x_* - x_0) < x_0 + x_* - x_0 < x_*.
\end{aligned}$$

To prove the second conclusion, recall that $\alpha(x_* - x)^2 \leq -\nabla f(x)(x_* - x) \leq \beta(x_* - x)^2$, thus for $x \in (x_0, x_*)$, we obtain $\alpha(x_* - x) \leq -\nabla f(x) \leq \beta(x_* - x)$.

As for the third conclusion, since $\alpha(x - x_0)^2 \leq (\nabla f(x) - \nabla f(x_0))(x - x_0) \leq \beta(x - x_0)^2$, thus for $x \in (x_0, x_*)$, we obtain $\alpha(x - x_0) + \nabla f(x_0) \leq \nabla f(x) \leq \beta(x - x_0) + \nabla f(x_0)$. which completes our proof. ☐

C. Missing proofs in main text

C.1. Proof of Theorem 1

Proof. The first part of the theorem is an extension of Proposition 1 and Proposition 2 in (Neu & Rosasco, 2018). Beyond the analysis of constant learning rate in (Neu & Rosasco, 2018), we show the corresponding results for adaptive learning rates.

Recall the SGD updates for linear regression problem

$$w_{k+1} = w_k - \eta_k(x_{k+1}x_{k+1}^\top w_k - x_{k+1}y_{k+1}), \quad w_0 = 0.$$

Let

$$\Sigma = \mathbb{E}_x[xx^\top], \quad a = \mathbb{E}_{x,y}[xy], \quad w_* = \Sigma^{-1}a, \quad \epsilon_k = (\Sigma w_k - a) - (x_{k+1}x_{k+1}^\top w_k - x_{k+1}y_{k+1}),$$

where ϵ_k is the gradient noise, and $\mathbb{E}_{k+1}[\epsilon_k] = 0$. Under these notations we have

$$w_{k+1} = w_k - \eta_k(\Sigma w_k - a) + \eta_k\epsilon_k = w_k - \eta_k\Sigma(w_k - w_*) + \eta_k\epsilon_k, \quad w_0 = 0. \tag{12}$$

Similarly for linear regression with ℓ_2-regularization, SGD takes update

$$\hat{w}_{k+1} = \hat{w}_k - \gamma_k(x_{k+1}x_{k+1}^\top\hat{w}_k - x_{k+1}y_{k+1} + \lambda\hat{w}_k), \quad \hat{w}_0 = 0.$$

Let

$$\hat{w}_* = (\Sigma + \lambda I)^{-1}a,$$

then

$$\hat{w}_{k+1} = \hat{w}_k - \gamma_k(\Sigma\hat{w}_k - a + \lambda\hat{w}_k) + \gamma_k\epsilon_k = \hat{w}_k - \gamma_k(\Sigma + \lambda I)(\hat{w}_k - \hat{w}_*) + \gamma_k\epsilon_k, \quad \hat{w}_0 = 0. \tag{13}$$

Expectations First let us compute the expectations. For Eq. (12), after taking expectation at time $k+1$, we have

$$\mathbb{E}_{k+1}[w_{k+1}] = w_k - \eta_k \Sigma (w_k - w_*).$$

Then recursively taking expectation at time $k, \dots, 1$, we obtain

$$\mathbb{E}[w_{k+1}] = \mathbb{E}[w_k] - \eta_k \Sigma (\mathbb{E}[w_k] - w_*), \quad \mathbb{E}[w_0] = w_0 = 0.$$

Solving the above recurrence relation we have

$$\mathbb{E}[w_k] - w_* = \Pi_{i=0}^{k-1} (I - \eta_i \Sigma)(w_0 - w_*), \quad w_0 = 0, \quad w_* = \Sigma^{-1} a,$$

hence

$$\mathbb{E}[w_{k+1}] - \mathbb{E}[w_k] = -\Pi_{i=0}^{k-1} (I - \eta_i \Sigma) \eta_k \Sigma (w_0 - w_*) = \Pi_{i=0}^{k-1} (I - \eta_i \Sigma) \eta_k a, \quad \mathbb{E}[w_0] = 0.$$

In a same way we can solve Eq. (13) in expectation and obtain

$$\mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k] = \Pi_{i=0}^{k-1} (I - \gamma_i (\Sigma + \lambda I)) \gamma_k a, \quad \mathbb{E}[\hat{w}_0] = 0.$$

Notice that the weighting scheme is defined by

$$P_k = 1 - \Pi_{i=0}^{k} (1 - \lambda \gamma_i),$$

and $1 - \lambda \gamma_i = \frac{\gamma_i}{\eta_i}$, we can directly verify that

$$\mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k] = (1 - P_k)(\mathbb{E}[w_{k+1}] - \mathbb{E}[w_k]).$$

Thus by Lemma 2, we know that

$$P_k \mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1 - P_k)\mathbb{E}[w_k], \quad k \geq 0.$$

Hence the first conclusion holds.

Convergence By assumptions we know $0 < \eta \leq \eta_i < \frac{1}{\beta} \leq \frac{1}{\lambda_{\max}}$, where $\lambda_{\max}$ is the largest eigenvalue of Σ. Thus

$$\left\| \mathbb{E}[w_k] - w_* \right\|_2 \leq \left\| \Pi_{i=0}^{k-1} (I - \eta_i \Sigma) \right\|_2 \cdot \left\| w_0 - w_* \right\|_2 \leq \left\| (I - \eta \Sigma)^k \right\|_2 \cdot \left\| w_0 - w_* \right\|_2 \to 0,$$

and $\lim_{k \to +\infty} \mathbb{E}[w_k] = w_* = \Sigma^{-1} a$.

In a similar manner, since $\gamma_i = \frac{\eta_i}{1 + \eta_i \lambda}$ and $0 < \eta \leq \eta_i < \frac{1}{\beta} \leq \frac{1}{\lambda_{\max}}$, we have $0 < \frac{\eta}{1 + \lambda \eta} = \gamma \leq \gamma_i < \frac{1}{\beta + \lambda} \leq \frac{1}{\lambda_{\max} + \lambda}$. Thus

$$\left\| \mathbb{E}[\hat{w}_k] - \hat{w}_* \right\|_2 \leq \left\| \Pi_{i=0}^{k-1} (I - \gamma_i (\Sigma + \lambda I)) \right\|_2 \cdot \left\| \hat{w}_0 - \hat{w}_* \right\|_2 \leq \left\| (I - \gamma (\Sigma + \lambda I))^k \right\|_2 \cdot \left\| \hat{w}_0 - \hat{w}_* \right\|_2 \to 0,$$

and $\lim_{k \to +\infty} \mathbb{E}[\hat{w}_k] = \hat{w}_* = (\Sigma + \lambda I)^{-1} a$.

On the other hand, by the first conclusion we know

$$\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] = (1 - P_k)(\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]).$$

Since $\mathbb{E}[w_k]$ converges, $\mathbb{E}[\tilde{w}_k] = P_k^{-1} \sum_{i=1}^{k} p_i \mathbb{E}[w_i]$ is bounded. Therefore

$$\left\| \mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] \right\|_2 = (1 - P_k) \left\| \mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k] \right\|_2 = \mathcal{O}(1 - P_k) = \mathcal{O}(\Pi_{i=0}^{k} (1 - \lambda \gamma_i)) \leq \mathcal{O}((1 - \lambda \gamma)^k).$$

Hence the second claim is true.

Variance Now we turn to analyze the deviation of the averaged solution. From Eq. (12), we can recursively obtain

$$w_i = \mathbb{E}[w_i] + \xi_i, \quad \xi_i = \sum_{j=0}^{i-1} \Pi_{h=j+1}^{i-1}(I - \eta_h \Sigma)\eta_j \epsilon_j,$$

where we abuse the notation and let $\Pi_{h=i}^{i-1}(I - \eta_h \Sigma) = I$.

Now applying iterate averaging with respect to $p_i = \lambda\gamma_i \Pi_{h=0}^{i-1}(1 - \lambda\gamma_h)$, we have

$$P_k \tilde{w}_k = \sum_{i=1}^{k} p_i w_i = \sum_{i=1}^{k} p_i \mathbb{E}[w_i] + \sum_{i=1}^{k} p_i \xi_i = P_k \mathbb{E}[\tilde{w}_k] + \sum_{i=1}^{k} p_i \xi_i.$$

We turn to calculate the noise term $\sum_{i=1}^{k} p_i \xi_i$. Note that in every step, all of the matrices can be diagonalized simultaneously, thus they commute, similarly hereinafter.

$$\begin{aligned}
&\sum_{i=1}^{k} p_i \xi_i = \sum_{i=1}^{k} p_i \left(\sum_{j=0}^{i-1} \Pi_{h=j+1}^{i-1}(I - \eta_h \Sigma)\eta_j \epsilon_j \right) \\
=& \sum_{j=0}^{k-1} \left(\sum_{i=j+1}^{k} p_i \Pi_{h=j+1}^{i-1}(I - \eta_h \Sigma)\eta_j \right) \epsilon_j \\
=& \sum_{j=0}^{k-1} \left(\sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=0}^{i-1}(1 - \lambda\gamma_h) \Pi_{h=j+1}^{i-1}(I - \eta_h \Sigma)\eta_j \right) \epsilon_j \\
=& \sum_{j=0}^{k-1} \left(\sum_{i=j+1}^{k} \lambda\gamma_i \left(\Pi_{h=0}^{j-1}(1 - \lambda\gamma_h) \right) \left(\Pi_{h=j+1}^{i-1}(1 - \lambda\gamma_h)(I - \eta_h \Sigma) \right) ((1 - \lambda\gamma_j)\eta_j) \right) \epsilon_j \\
=& \sum_{j=0}^{k-1} \left(\left(\Pi_{h=0}^{j-1}(1 - \lambda\gamma_h) \right) \left(\sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=j+1}^{i-1} \left(I - \gamma_h(\Sigma + \lambda I) \right) \right) \gamma_j \right) \epsilon_j \\
=& \sum_{j=0}^{k-1} A_j \epsilon_j,
\end{aligned}$$

where $A_j = \gamma_j \left(\Pi_{h=0}^{j-1}(1 - \lambda\gamma_h) \right) \left(\sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=j+1}^{i-1} \left(I - \gamma_h(\Sigma + \lambda I) \right) \right)$. Recall that $\epsilon_0, \epsilon_1 \dots, \epsilon_k$ is a martingale difference sequence, then $\sum_{i=1}^{k} p_i \xi_i = \sum_{j=0}^{k-1} A_j \epsilon_j$ is a martingale. Thus

$$\mathrm{Tr}\,\mathrm{Var}\left[\sum_{i=1}^{k} p_i \xi_i \right] = \mathrm{Tr}\,\mathrm{Var}\left[\sum_{j=0}^{k-1} A_j \epsilon_j \right] = \sum_{j=0}^{k-1} \mathrm{Tr}\,\mathrm{Var}\left[A_j \epsilon_j \right],$$

where “Var” is the covariance of a random vector. and “Tr” is the trace of a matrix.

Next we bound each term in the summation as

$$\mathrm{Tr}\,\mathrm{Var}\left[A_j \epsilon_j \right] = \mathrm{Tr}\,\mathbb{E}\left[(A_j \epsilon_j)(A_j \epsilon_j)^\top \right] = \mathbb{E}\left[\|A_j \epsilon\|_2^2 \right] \le \|A_j\|_2^2 \cdot \mathbb{E}\left[\|\epsilon\|_2^2 \right] \le \sigma^2 \|A_j\|_2^2 .$$

And we remain to bound $\|A_j\|_2^2$. Remember that $\eta \le \eta_h \le \frac{1}{\beta},\ \gamma \le \gamma_h \le \frac{1}{\lambda+\beta}$, we have

$$\begin{aligned}
\|A_j\|_2^2 =& \left\|\gamma_j \left(\Pi_{h=0}^{j-1}(1-\lambda\gamma_h)\right)\left(\sum_{i=j+1}^{k}\lambda\gamma_i\Pi_{h=j+1}^{i-1}\left(I-\gamma_h(\Sigma+\lambda I)\right)\right)\right\|_2^2\\
\le& \left\|\frac{1}{\lambda+\beta}\left((1-\lambda\gamma)^j\right)\left(\sum_{i=j+1}^{k}\frac{\lambda}{\lambda+\beta}\left(I-\gamma(\Sigma+\lambda I)\right)^{i-j-1}\right)\right\|_2^2\\
=& \left\|\frac{\lambda}{(\lambda+\beta)^2}\left((1-\lambda\gamma)^j\right)\left(\sum_{i=0}^{k-j-1}\left(I-\gamma(\Sigma+\lambda I)\right)^{i}\right)\right\|_2^2\\
\le& \left(\frac{\lambda}{(\lambda+\beta)^2}\left((1-\lambda\gamma)^j\right)\left(\sum_{i=0}^{k-j-1}\left(1-\gamma(\alpha+\lambda)\right)^{i}\right)\right)^2\\
\le& \left(\frac{\lambda}{(\lambda+\beta)^2}\left((1-\lambda\gamma)^j\right)\left(\frac{1}{\gamma(\alpha+\lambda)}\right)\right)^2\\
=& \frac{\lambda^2}{\gamma^2(\lambda+\alpha)^2(\lambda+\beta)^4}(1-\lambda\gamma)^{2j}.
\end{aligned}$$

The second equality holds because $\alpha \le \lambda_{min}(\Sigma)$.

Based on previous discussion we have

$$\begin{aligned}
&\operatorname{Tr}\operatorname{Var}\left[\sum_{i=1}^{k}p_i\xi_i\right] = \sum_{j=0}^{k-1}\operatorname{Tr}\operatorname{Var}\left[A_j\epsilon_j\right] \le \sum_{j=0}^{k-1}\sigma^2\|A_j\|_2^2\\
\le& \sum_{j=0}^{k-1}\frac{\lambda^2\sigma^2}{\gamma^2(\lambda+\alpha)^2(\lambda+\beta)^4}(1-\lambda\gamma)^{2j} \le \frac{\lambda^2\sigma^2}{\gamma^2(\lambda+\alpha)^2(\lambda+\beta)^4}\frac{1}{1-(1-\lambda\gamma)^2}\\
=& \frac{\lambda\sigma^2}{\gamma^3(2-\lambda\gamma)(\lambda+\alpha)^2(\lambda+\beta)^4}.
\end{aligned}$$

Now by multivariate Chebyshev's inequality, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{k}p_i\xi_i\right\|_2 \ge \epsilon\right) \le \frac{\operatorname{Tr}\operatorname{Var}\left[\sum_{i=1}^{k}p_i\xi_i\right]}{\epsilon^2} \le \frac{\lambda\sigma^2}{\epsilon^2\gamma^3(2-\lambda\gamma)(\lambda+\alpha)^2(\lambda+\beta)^4} = \delta.$$

That is, with probability at least $1-\delta$, we have

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 = \left\|\sum_{i=1}^{k}p_i\xi_i\right\|_2 \le \epsilon,$$

where

$$\epsilon = \frac{\sigma}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}.$$

This completes our proof. ☐

C.2. Proof of Theorem 1.1

Proof. The derivation of kernel ridge regression can be found in (Mohri et al., 2018). We consider the following loss function of the dual problem

$$L(\alpha,\lambda) = \frac{1}{2}\|y-K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha,$$

where $y = (y_1, \ldots, y_n)^T$ is the label set. Then GD takes update

$$\alpha_{k+1} = \alpha_k - \eta_k \left(K^2 \alpha_k - Ky + \lambda K \alpha_k \right), \quad \alpha_0 = 0.$$

Let $\alpha_* = (K + \lambda I)^{-1} y$, then

$$\alpha_{k+1} - \alpha_* = \left(I - \eta_k (K^2 + \lambda K) \right) (\alpha_k - \alpha_*),$$

thus

$$\alpha_{k+1} - \alpha_* = \Pi_{i=0}^{k} \left(I - \eta_i (K^2 + \lambda K) \right) (\alpha_0 - \alpha_*),$$

and

$$\alpha_{k+1} - \alpha_k = \Pi_{i=0}^{k-1} \left(I - \eta_i (K^2 + \lambda K) \right) \cdot \eta_k (K^2 + \lambda K) \cdot (K + \lambda I)^{-1} y = \Pi_{i=0}^{k-1} \left(I - \eta_i (K^2 + \lambda K) \right) \eta_k K y.$$

Similarly for $\hat{\alpha}_k$, i.e., the GD path for $L(\hat{\alpha}, \hat{\lambda})$ with learning rate γ_k, we have

$$\hat{\alpha}_{k+1} - \hat{\alpha}_k = \Pi_{i=0}^{k-1} \left(I - \gamma_i (K^2 + \hat{\lambda} K) \right) \gamma_k K y.$$

We emphasize that the generalized learning rate $\gamma_k = \left(I + (\hat{\lambda} - \lambda) \eta_k K \right)^{-1} \eta_k$ commutes with K. And

$$I - \gamma_k (\hat{\lambda} - \lambda) K = \frac{\gamma_k}{\eta_k}.$$

Thus for the generalized weighting scheme $P_K = 1 - \Pi_{i=0}^{k} (\gamma_i / \eta_i)$ we have

$$\begin{aligned}
(1 - P_k)(\alpha_{k+1} - \alpha_k) &= \Pi_{i=0}^{k-1} \left(\frac{\gamma_i}{\eta_i} \left(I - \eta_i (K^2 + \lambda K) \right) \right) \frac{\gamma_k}{\eta_k} \eta_k K y \\
&= \Pi_{i=0}^{k-1} \left(\frac{\gamma_i}{\eta_i} - \gamma_i (K^2 + \lambda K) \right) \gamma_k K y = \Pi_{i=0}^{k-1} \left(I - \gamma_i (\hat{\lambda} - \lambda) K - \gamma_i (K^2 + \lambda K) \right) \gamma_k K y \\
&= \Pi_{i=0}^{k-1} \left(I - \gamma_i (K^2 + \hat{\lambda} K) \right) \gamma_k K y = \hat{\alpha}_{k+1} - \hat{\alpha}_k.
\end{aligned}$$

Therefore by Lemma 2 we have

$$P_k \tilde{\alpha}_k = \hat{\alpha}_k - (1 - P_k) \alpha_k.$$

Let $\lambda_{\max}$ and $\lambda_{\min}$ be the maximal and minimal eigenvalue of K respectively. Then if

$$\eta \leq \eta_k \leq \max \left\{ \frac{1}{\lambda_{\max}(\lambda_{\max} + \lambda)}, \frac{1}{\lambda_{\max}(\lambda_{\max} + 2\hat{\lambda} - \lambda)} \right\}, \quad \gamma = \left(I + (\hat{\lambda} - \lambda) \eta K \right)^{-1} \eta,$$

we have

$$\eta (K^2 + \lambda K) \preceq \eta_k (K^2 + \lambda K) \prec I, \quad \gamma (K^2 + \hat{\lambda} K) \preceq \gamma_k (K^2 + \hat{\lambda} K) \prec I,$$

which guarantees the convergence of α_k and $\hat{\alpha}_k$. Hence both α_k and $\tilde{\alpha}_k$ are bounded. And the convergence rate is given by

$$\|\hat{\alpha}_k - \tilde{\alpha}_k\|_2 = \|(1 - P_k)(\alpha_k - \tilde{\alpha}_k)\|_2 = \mathcal{O}(\|1 - P_k\|_2) \leq \mathcal{O}\left(\|\gamma / \eta\|_2^k \right) = \mathcal{O}\left((1 + (\hat{\lambda} - \lambda) \eta \lambda_{\min})^{-k} \right).$$

□

C.3. Proof of Theorem 2

Proof. Let us consider changing of variable $v_k = Q^{\frac{1}{2}} w_k$, then

$$\begin{aligned}
v_{k+1} &= Q^{\frac{1}{2}} w_{k+1} = Q^{\frac{1}{2}} w_k - \eta_k Q^{-\frac{1}{2}} (x_k x_k^\top w_k - x_k y_k) \\
&= Q^{\frac{1}{2}} w_k - \eta_k (Q^{-\frac{1}{2}} x_k x_k^T Q^{-\frac{1}{2}} Q^{\frac{1}{2}} w_k - Q^{-\frac{1}{2}} x_k y_k) \\
&= v_k - \eta_k (Q^{-\frac{1}{2}} x_k x_k^\top Q^{-\frac{1}{2}} v_k - Q^{-\frac{1}{2}} x_k y_k).
\end{aligned}$$

Similarly let $\hat{v}_k = Q^{\frac{1}{2}}\hat{w}_k$, then

$$\begin{aligned}\hat{v}_{k+1} = Q^{\frac{1}{2}}\hat{w}_{k+1} = Q^{\frac{1}{2}}\hat{w}_k - \gamma_k Q^{-\frac{1}{2}}(x_k x_k^\top \hat{w}_k - x_k y_k - \lambda Q \hat{w}_k)\\ =& Q^{\frac{1}{2}}\hat{w}_k - \gamma_k (Q^{-\frac{1}{2}} x_k x_k^\top Q^{-\frac{1}{2}} Q^{\frac{1}{2}}\hat{w}_k - Q^{-\frac{1}{2}} x_k y_k - \lambda Q^{\frac{1}{2}}\hat{w}_k)\\ =& \hat{v}_k - \gamma_k (Q^{-\frac{1}{2}} x_k x_k^\top Q^{-\frac{1}{2}}\hat{v}_k - Q^{-\frac{1}{2}} x_k y_k - \lambda \hat{v}_k).\end{aligned}$$

Let us denote

$$\Sigma = \mathbb{E}_x[xx^T], \quad a = \mathbb{E}_{x,y}[xy], \quad w_* = \Sigma^{-1}a, \quad \hat{w}_* = (\Sigma + \lambda I)^{-1}a, \quad \epsilon_k = (\Sigma w_k - a) - (x_{k+1}x_{k+1}^\top w_k - x_{k+1}y_{k+1}),$$

and correspondingly,

$$\Lambda = Q^{-\frac{1}{2}}\Sigma Q^{-\frac{1}{2}}, \quad b = Q^{-\frac{1}{2}}a, \quad v_* = Q^{-\frac{1}{2}}w_*, \quad \hat{v}_* = Q^{-\frac{1}{2}}\hat{w}_*, \quad \iota_k = Q^{-\frac{1}{2}}\epsilon_k.$$

Under these notations we have

$$v_{k+1} = v_k - \eta_k(\Lambda v_k - b) + \eta_k \iota_k, \quad v_0 = 0. \tag{14}$$

and

$$\hat{v}_{k+1} = \hat{v}_k - \gamma_k(\Lambda \hat{v}_k - b + \lambda \hat{v}_k) + \gamma_k \iota_k, \quad \hat{v}_0 = 0. \tag{15}$$

We can see that Eq. (14) and Eq. (15) are exactly what we have studied in Theorem 1. Also by assumption we know

$$\alpha I \preceq \Lambda \preceq \beta I.$$

Thus by Theorem 1 we have the following conclusions:

1. In expectation for any $k > 0$,
$$P_k \mathbb{E}[\tilde{v}_k] = \mathbb{E}[\hat{v}_k] - (1 - P_k)\mathbb{E}[v_k].$$
2. Both $\mathbb{E}[v_k]$ and $\mathbb{E}[\hat{v}_k]$ converge. And there exists a constant K such that for all $k > K$,
$$\|\mathbb{E}[\hat{v}_k] - \mathbb{E}[\tilde{v}_k]\|_2 \le \mathcal{O}((1-\lambda\gamma)^k).$$
Hence the limitation of $\mathbb{E}[\tilde{v}_k]$ exists and $\lim_{k\to\infty}\mathbb{E}[\tilde{v}_k] = \lim_{k\to\infty}\mathbb{E}[\hat{v}_k]$.
3. If the noise ι_k has uniform bounded variance
$$\mathbb{E}[\|\tilde{\iota}_k\|_2^2] \le \|Q\|_2 \sigma^2, \quad \forall k.$$
Then for k large enough, with probability at least $1-\delta$, we have
$$\|P_k \tilde{v}_k - P_k \mathbb{E}[\tilde{v}_k]\|_2 \le \epsilon,$$
where
$$\epsilon = \frac{\|Q\|_2^{\frac{1}{2}}\sigma}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}.$$

Now let $w_k = Q^{-\frac{1}{2}}v_k$, $\hat{w}_k = Q^{-\frac{1}{2}}\hat{v}_k$, then $\tilde{w}_k = \frac{1}{P_k}\sum_{i=1}^k p_i w_i = Q^{-\frac{1}{2}}\frac{1}{P_k}\sum_{i=1}^k p_i v_i = Q^{-\frac{1}{2}}\tilde{v}_k$. Hence we have

1. In expectation for any $k > 0$,
$$P_k \mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1 - P_k)\mathbb{E}[w_k].$$
2. Both $\mathbb{E}[w_k]$ and $\mathbb{E}[\hat{w}_k]$ converge. And there exists a constant K such that for all $k > K$,
$$\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 \le \mathcal{O}((1-\lambda\gamma)^k).$$
Hence the limitation of $\mathbb{E}[\tilde{w}_k]$ exists and $\lim_{k\to\infty}\mathbb{E}[\tilde{w}_k] = \lim_{k\to\infty}\mathbb{E}[\hat{w}_k]$.

3. If the PSGD noise $Q^{-1}\epsilon_k$ has uniform bounded variance

$$\mathbb{E}[\left\|Q^{-1}\epsilon_i\right\|_2^2] \le \sigma^2, \quad \forall i.$$

Then for k large enough, with probability at least $1-\delta$, we have

$$\|P_k \tilde{w}_k - P_k \mathbb{E}[\tilde{w}_k]\|_2 \le \epsilon,$$

where

$$\epsilon = \frac{\sigma \left\|Q^{-\frac{1}{2}}\right\|_2 \cdot \left\|Q^{\frac{1}{2}}\right\|_2}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}} \le \frac{\sigma \left\|Q\right\|_2}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}.$$

Hence our claims are proved.

□

C.4. Proof of Theorem 3

Proof. First, provided $0 < \eta < \frac{1}{\beta} < \frac{1}{\alpha}$ and $\gamma = \frac{1}{\frac{1}{\eta}+\lambda}$, we have

$$\frac{\eta\alpha}{\alpha+\lambda} = \frac{1}{\frac{1}{\eta}+\frac{\lambda}{\eta\alpha}} < \frac{1}{\frac{1}{\eta}+\lambda} = \gamma < \frac{1}{\beta+\lambda} \le \frac{1}{\alpha+\lambda}.$$

Therefore $0 < \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} < 1$, and

$$P_k = 1 - \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad p_k = P_k - P_{k-1},$$

is a well defined weighting scheme, i.e., P_k is non-negative, non-decreasing and $\lim_{k\to\infty} P_k = 1$.

Recall the NSGD updates for linear regression problem

$$w_{k+1} = v_k - \eta(x_{k+1}x_{k+1}^\top v_k - x_{k+1}y_{k+1}), \quad v_k = w_k + \tau(w_k - w_{k-1}), \quad w_0 = w_1 = 0,$$

where $\tau = \frac{1-\sqrt{\eta\alpha}}{1+\sqrt{\eta\alpha}}$.

Let

$$\Sigma = \mathbb{E}_x[xx^\top], \quad a = \mathbb{E}_{x,y}[xy], \quad \epsilon_k = (\Sigma v_k - a) - (x_{k+1}x_{k+1}^\top v_k - x_{k+1}y_{k+1}),$$

where ϵ_k is the gradient noise, and $\mathbb{E}_{k+1}[\epsilon_k] = 0$. Under these notations we have

$$w_{k+1} = v_k - \eta(\Sigma v_k - a) + \eta\epsilon_k, \quad v_k = w_k + \tau(w_k - w_{k-1}), \quad w_0 = w_1 = 0.$$

Thus

$$w_{k+1} = (1+\tau)(1-\eta\Sigma)w_k - \tau(1-\eta\Sigma)w_{k-1} + \eta a + \eta\epsilon_k, \quad w_0 = w_1 = 0. \tag{16}$$

Similarly for the linear regression with ℓ_2-regularization, NSGD takes update

$$\hat{w}_{k+1} = \hat{v}_k - \gamma\left((x_{k+1}x_{k+1}^T + \lambda)\hat{v}_k - x_{k+1}y_{k+1}\right), \quad \hat{v}_k = \hat{w}_k + \hat{\tau}(\hat{w}_k - \hat{w}_{k-1}), \quad \hat{w}_0 = \hat{w}_1 = 0,$$

where $\hat{\tau} = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1+\sqrt{\gamma(\alpha+\lambda)}}$.

And we have

$$\hat{w}_{k+1} = (1+\hat{\tau})\left(1-\gamma(\Sigma+\lambda)\right)\hat{w}_k - \hat{\tau}\left(1-\gamma(\Sigma+\lambda)\right)\hat{w}_{k-1} + \gamma a + \gamma\epsilon_k, \quad \hat{w}_0 = \hat{w}_1 = 0. \tag{17}$$

Expectation First let us compute the expectations. Let $z_k = \mathbb{E}[w_{k+1}] - \mathbb{E}[w_k],\ \hat{z}_k = \mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k]$, we aim to show that

$$(1 - P_k) z_k = \hat{z}_k, \quad k \geq 0. \tag{18}$$

Then according to Lemma 2, we prove the first conclusion in Theorem 3.

We begin with solving z_k.

For Eq. (16), taking expectation with respect to the random mini-batch sampling procedure, we have

$$\mathbb{E}[w_{k+1}] = (1+\tau)(1-\eta\Sigma)\mathbb{E}[w_k] - \tau(1-\eta\Sigma)\mathbb{E}[w_{k-1}] + \eta a, \quad \mathbb{E}[w_0] = \mathbb{E}[w_1] = 0.$$

Thus $z_k = \mathbb{E}[w_{k+1}] - \mathbb{E}[w_k]$ satisfies

$$z_{k+1} = (1+\tau)(1-\eta\Sigma) z_k - \tau(1-\eta\Sigma) z_{k-1}, \quad z_0 = 0, \quad z_1 = \eta a. \tag{19}$$

Without loss of generality, let us assume Σ is diagonal in the following. Otherwise consider its eigenvalue decomposition $\Sigma = U\Lambda U^T$, and replace z_k with $U^\top z_k$. All of the operators in the following are defined entry-wisely.

Eq. (19) defines a homogeneous linear recurrence relation with constant coefficients, which could be solved in a standard manner. Let

$$A = (1+\tau)(1-\eta\Sigma) = \frac{2(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad B = -\tau(1-\eta\Sigma) = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}},$$

then the characteristic function of Eq. (19) is

$$r^2 - Ar - B = 0. \tag{20}$$

Since Σ is diagonal, $0 < \eta < \frac{1}{\alpha}$, and α is no greater than the smallest eigenvalue of Σ, we have

$$A^2 + 4B = \frac{4\eta(1-\eta\Sigma)(\alpha - \Sigma)}{(1+\sqrt{\eta\alpha})^2} \leq 0.$$

Thus the characteristic function (20) has two conjugate complex roots r_1 and r_2 (they might be equal). Suppose $r_{1,2} = s \pm ti$. Then the solution of Eq. (19) can be written as

$$z_k = 2(-B)^{\frac{k}{2}} \left(E\cos(\theta k) + F\sin(\theta k) \right), \quad k \geq 0,$$

where E and F are constants decided by initial conditions $z_0 = 0,\ z_1 = \eta a$, and θ satisfies

$$\cos\theta = \frac{s}{\sqrt{s^2+t^2}}, \quad \sin\theta = \frac{t}{\sqrt{s^2+t^2}}, \quad r_{1,2} = s \pm ti.$$

Since $2s = r_1 + r_2 = A$, $s^2 + t^2 = r_1 \dot{r}_2 = -B$, we have

$$\cos\theta = \frac{A}{2\sqrt{-B}} = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin\theta = \frac{\sqrt{-4B - A^2}}{2\sqrt{-B}} = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}.$$

Because $z_0 = 0, z_1 = \eta a$, we know that

$$E = 0, \quad 2F = \frac{\eta a}{(-B)^{\frac{1}{2}}\sin\theta}.$$

Thus

$$z_k = \frac{\eta a}{\sin\theta}(-B)^{\frac{k-1}{2}}\sin(\theta k), \quad k \geq 0. \tag{21}$$

where

$$B = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad \cos\theta = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin\theta = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}.$$

One can directly verify that Eq. (21) solves the recurrence relation (19).

Then we solve $\hat{z}_k$.

Similarly treat Eq. (17), we know $\hat{z}_k = \mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k]$ satisfies

$$\hat{z}_{k+1} - (1+\hat{\tau})\left(1-\gamma(\Sigma+\lambda)\right)\hat{z}_k + \hat{\tau}\left(1-\gamma(\Sigma+\lambda)\right)\hat{z}_{k-1} = 0, \quad \hat{z}_0 = 0, \quad \hat{z}_1 = -\gamma a.$$

Repeat the calculation, we obtain

$$\hat{z}_k = \frac{\gamma a}{\sin\hat{\theta}}(-\hat{B})^{\frac{k-1}{2}}\sin(\hat{\theta}k), \quad k \ge 0,$$

where

$$\hat{B} = \frac{-\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)\left(1-\gamma(\Sigma+\lambda)\right)}{1+\sqrt{\gamma(\alpha+\lambda)}},$$
$$\cos\hat{\theta} = \sqrt{\frac{1-\gamma(\Sigma+\lambda)}{1-\gamma(\alpha+\lambda)}}, \quad \sin\hat{\theta} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{1-\gamma(\alpha+\lambda)}}.$$

Finally we verify the sufficient condition in Lemma 2 (Eq. (18)).

First we show that if $1-\lambda\gamma = \frac{\gamma}{\eta}$, we have $\theta \equiv \hat{\theta} \pmod{2\pi}$. To see this, we only need to verify that $\cos\hat{\theta} = \cos\theta,\ \sin\hat{\theta} = \sin\theta$. This is because

$$\cos\hat{\theta} = \sqrt{\frac{1-\gamma\lambda-\gamma\Sigma}{1-\gamma\lambda-\gamma\alpha}} = \sqrt{\frac{\frac{\gamma}{\eta}-\gamma\Sigma}{\frac{\gamma}{\eta}-\gamma\alpha}} = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}} = \cos\theta;$$
$$\sin\hat{\theta} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{1-\gamma\lambda-\gamma\alpha}} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{\frac{\gamma}{\eta}-\gamma\alpha}} = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}} = \sin\theta.$$

Therefore we have

$$z_k = \frac{\eta a}{\sin\theta}(-B)^{\frac{k-1}{2}}\sin(\theta k), \quad \hat{z}_k = \frac{\gamma a}{\sin\theta}(-\hat{B})^{\frac{k-1}{2}}\sin(\theta k).$$

Since

$$1-P_k = \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad \frac{\gamma}{\eta} = 1-\lambda\gamma,$$

we have

$$\frac{\eta}{\gamma}(1-P_k)(-B)^{\frac{k-1}{2}} = \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2}{(1-\sqrt{\eta\alpha})^2}\cdot\frac{(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}\right)^{\frac{k-1}{2}}$$
$$= \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2(1-\eta\Sigma)}{1-\eta\alpha}\right)^{\frac{k-1}{2}} = \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2\left(1-\gamma(\Sigma+\lambda)\right)}{1-\gamma(\alpha+\lambda)}\right)^{\frac{k-1}{2}}$$
$$= \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)\left(1-\gamma(\Sigma+\lambda)\right)}{1+\sqrt{\gamma(\alpha+\lambda)}}\right)^{\frac{k-1}{2}} = (-\hat{B})^{\frac{k-1}{2}}.$$

Thus $(1-P_k)z_k = \hat{z}_k$. And according to Lemma 2, we have

$$\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] = (1-P_k)\left(\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]\right), \quad k \ge 0.$$

Hence the first conclusion holds.

Convergence Since $L(w)$ is β-smooth, and the corresponding learning rate $\eta < \frac{1}{\beta}$, $\mathbb{E}[w_k]$ converges (Beck & Teboulle, 2009). Similarly, $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_2^2$ is $(\beta+\lambda)$-smooth, and the corresponding learning rate $\gamma = \frac{1}{\frac{1}{\eta}+\lambda} < \frac{1}{\beta+\lambda}$, thus $\mathbb{E}[\hat{w}_k]$ converges (Beck & Teboulle, 2009). Specially for linear regression, these can be also verified by noticing that $0 < -B < 1$ because $\eta < \frac{1}{\beta}$ and

$$\sum_{i=1}^{k} |z_i| = \sum_{i=1}^{k} \left| \frac{\eta a}{\sin\theta} (-B)^{\frac{i-1}{2}} \sin(\theta i) \right| \leq \sum_{i=1}^{k} \left| \frac{\eta a}{\sin\theta} (-B)^{\frac{i-1}{2}} \right| < +\infty,$$

i.e., the right hand side of the above series converge, which implies that $\mathbb{E}[w_k] = \sum_{i=1}^{k} z_i$ converges absolutely, hence it converges. In a same manner $\mathbb{E}[\hat{w}_k]$ converges. Thus there exist constants M and K such that for all $k > K$, $\|\mathbb{E}[w_k]\|_2 \leq M$, $\|\mathbb{E}[\hat{w}_k]\|_2 \leq M$. Hence

$$\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 = (1 - P_k) \|\mathbb{E}[w_k] - \mathbb{E}[\hat{w}_k]\|_2 \leq \frac{\gamma}{\eta} C^{k-1} \cdot 2M = \mathcal{O}(C^k),$$

where $C = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \in (0, 1)$, thus by taking limitation in both sides we obtain

$$\lim_{k\to\infty} \mathbb{E}[\tilde{w}_k] = \lim_{k\to\infty} \mathbb{E}[\hat{w}_k],$$

Hence the second conclusion holds.

Variance Next we turn to analyze the deviation of the averaged solution.

Let $w_i = \mathbb{E}[w_i] + \xi_i$. Based on Eq. (16), we first prove that

$$\xi_i = \sum_{j=1}^{i-1} a_{i-j} \eta \epsilon_j, \quad i \geq 1, \tag{22}$$

where

$$a_{k+1} = A a_k + B a_{k-1}, \quad a_0 = 0, \quad a_1 = 1. \tag{23}$$

We prove Eq. (22) by mathematical induction.

For $i = 1, 2$, by Eq. (16) we know $\xi_1 = w_1 - \mathbb{E}[w_1] = 0$ and $\xi_2 = w_2 - \mathbb{E}[w_2] = \eta\epsilon_1$, thus Eq. (22) holds. Now suppose Eq. (22) holds for $i - 1$ and i, then we consider $i + 1$. In Eq. (16), since $\xi_i = w_i - \mathbb{E}[w_i]$, taking difference we have

$$\xi_{i+1} = A\xi_i + B\xi_{i-1} + \eta\epsilon_i.$$

Now combining the induction assumptions we have

$$\begin{aligned}
\xi_{i+1} =& A \sum_{j=1}^{i-1} a_{i-j} \eta \epsilon_j + B \sum_{j=1}^{i-2} a_{i-j-1} \eta \epsilon_j + \eta \epsilon_i \\
=& \sum_{j=1}^{i-2} (A a_{i-j} + B a_{i-j-1}) \eta \epsilon_j + A a_1 \eta \epsilon_{i-1} + \eta \epsilon_i \\
=& \sum_{j=1}^{i-2} a_{i-j+1} \eta \epsilon_j + a_2 \eta \epsilon_{i-1} + a_1 \eta \epsilon_i \\
=& \sum_{j=1}^{i} a_{i-j+1} \eta \epsilon_j.
\end{aligned}$$

Thus by mathematical induction Eq. (22) is true for all $i \geq 1$.

Similarly to solve z_k, we can solve the recurrence relation Eq. (23) and obtain

$$a_k = \frac{1}{\sin\theta} (-B)^{\frac{k-1}{2}} \sin(\theta k), \quad k \geq 0, \tag{24}$$

where

$$B = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad \cos\theta = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin\theta = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}.$$

Thus

$$\begin{aligned}
\sqrt{-B} =& \sqrt{\frac{(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}} = (1-\sqrt{\eta\alpha})\sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \\
\frac{1}{\sin\theta} =& \sqrt{\frac{1-\eta\alpha}{\eta(\Sigma-\alpha)}} \preceq \sqrt{\frac{1-\eta\alpha}{\eta(\lambda_{\min}-\alpha)}} I,
\end{aligned}$$

where $\lambda_{\min}$ is the smallest eigenvalue of Σ.

Now apply iterate averaging with respect to

$$p_i = P_i - P_{i-1} = \frac{\gamma}{\eta}\left(\frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}}\right)\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{i-2},$$

we have

$$P_k\tilde{w}_k = \sum_{i=1}^k p_i w_i = \sum_{i=1}^k p_i \mathbb{E}[w_i] + \sum_{i=1}^k p_i\xi_i = P_k\mathbb{E}[\tilde{w}_k] + \sum_{i=1}^k p_i\xi_i.$$

We turn to calculate the noise term $\sum_{i=1}^k p_i\xi_i$. Note that in every step, all of the matrices can be diagonalized simultaneously, thus they commute, similarly hereinafter.

$$\begin{aligned}
\sum_{i=1}^k p_i\xi_i &= \sum_{i=1}^k p_i \sum_{j=1}^{i-1} a_{i-j}\eta\epsilon_j \\
&= \sum_{j=1}^{k-1}\left(\sum_{i=j+1}^k p_i a_{i-j}\right)\eta\epsilon_j \\
&= \sum_{j=1}^{k-1} A_j\epsilon_j,
\end{aligned}$$

where $A_j = \eta\sum_{i=j+1}^k p_i a_{i-j}$. Recall that $\epsilon_0, \epsilon_1 \ldots, \epsilon_k$ is a martingale difference sequence, $\sum_{i=1}^k p_i\xi_i = \sum_{j=0}^{k-1} A_j\epsilon_j$ is a martingale. Thus

$$\operatorname{Tr}\operatorname{Var}\left[\sum_{i=1}^k p_i\xi_i\right] = \operatorname{Tr}\operatorname{Var}\left[\sum_{j=1}^{k-1} A_j\epsilon_j\right] = \sum_{j=1}^{k-1}\operatorname{Tr}\operatorname{Var}\left[A_j\epsilon_j\right].$$

Next we bound each term in the summation as

$$\operatorname{Tr}\operatorname{Var}\left[A_j\epsilon_j\right] = \operatorname{Tr}\mathbb{E}\left[(A_j\epsilon_j)(A_j\epsilon_j)^\top\right] = \mathbb{E}\left[\|A_j\epsilon\|_2^2\right] \le \|A_j\|_2^2 \cdot \mathbb{E}\left[\|\epsilon\|_2^2\right] \le \sigma^2\|A_j\|_2^2.$$

And we remain to bound $\|A_j\|_2^2$:

$$
\begin{aligned}
\|A_j\|_2^2 &= \left\| \eta \sum_{i=j+1}^{k} p_i a_{i-j} \right\|_2^2 \\
&= \left\| \frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} \sum_{i=j+1}^{k} \left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{i-2} (-B)^{\frac{i-j-1}{2}} \sin(\theta(i-j)) \right\|_2^2 \\
&\leq \left\| \frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} \sum_{i=j+1}^{k} \left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{i-2} \left((1-\sqrt{\eta\alpha}) \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}} \right)^{i-j-1} \right\|_2^2 \\
&\leq \left(\frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} \sum_{i=j+1}^{k} \left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{i-2} (1-\sqrt{\eta\alpha})^{i-j-1} \right)^2 \\
&= \left(\frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} (1-\sqrt{\eta\alpha})^{1-j} \sum_{i=j+1}^{k} \left(1-\sqrt{\gamma(\alpha+\lambda)} \right)^{i-2} \right)^2 \\
&\leq \left(\gamma \sqrt{\frac{1-\eta\alpha}{\eta(\lambda_{\min}-\alpha)}} \cdot \frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{(1-\sqrt{\eta\alpha})^j} \cdot \frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^{j-1}}{\sqrt{\gamma(\alpha+\lambda)}} \right)^2 \\
&= \frac{\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{2j}.
\end{aligned}
$$

The first inequality is because $\sin(\theta(i-j)) \leq 1$, and the second inequality is because $\alpha < \lambda_{\min}(\Sigma)$.

Based on previous discussion we have

$$
\begin{aligned}
&\operatorname{Tr}\operatorname{Var}\left[\sum_{i=1}^{k} p_i \xi_i\right] = \sum_{j=1}^{k-1} \operatorname{Tr}\operatorname{Var}\left[A_j \epsilon_j\right] \leq \sum_{j=1}^{k-1} \sigma^2 \|A_j\|_2^2 \\
\leq & \sum_{j=1}^{k-1} \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{2j} \\
\leq & \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \cdot \frac{\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^2}{1-\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^2} \\
\leq & \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \cdot \frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2}{\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)} \\
= & \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)}.
\end{aligned}
$$

Now by multivariate Chebyshev's inequality, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{k} p_i \xi_i\right\|_2 \geq \epsilon\right) \leq \frac{\operatorname{Tr}\operatorname{Var}[\sum_{i=1}^{k} p_i\xi_i]}{\epsilon^2} \leq \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)}{\epsilon^2\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)} =: \delta.$$

That is, with probability at least $1-\delta$, we have

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 = \left\|\sum_{i=1}^{k} p_i\xi_i\right\|_2 \leq \epsilon,$$

where

$$\epsilon = \sqrt{\frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)}{\delta\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)}}.$$

This completes our proof.

□

C.5. Proof of Theorem 4

Proof. We will prove a stronger version of Theorem 4 by showing the conclusions hold for any 1-dim projection direction $v_1 \in \mathbb{R}^d$. Concisely, given a unit vector $v_1 \in \mathbb{R}^d$, we can extend it to a group of orthogonal basis, $v_1, v_2, \ldots, v_d$. For $w \in \mathbb{R}^d$, we denote its decomposition as

$$w = w^{(1)}v_1 + w^{(2)}v_2 + \cdots + w^{(d)}v_d, \quad w^{(i)} \in \mathbb{R}.$$

Define $h(w^{(1)}) = L(w) = L(w^{(1)}v_1 + \cdots + w^{(d)}v_d)$, then $\nabla h(w^{(1)}) = v_1^\top \nabla L(w)$. Now for one step of GD,

$$w_{k+1} = w_k - \eta\nabla L(w_k),$$

by multiplying v_1 in both sides, we obtain

$$w_{k+1}^{(1)} = v_1^\top w_{k+1} = v_1^\top w_k - \eta v_1^\top \nabla L(w_k) = w_k^{(1)} - \eta\nabla h(w_k^{(1)}). \tag{25}$$

We turn to study GD along direction v_1 by analyzing Eq. (25).

Firstly $h(w^{(1)})$ is α-strongly convex, β-smooth and lower bounded since $L(w)$ is α-strongly convex, β-smooth, and lower bounded. Let w_* be the unique minimum of $L(w)$, then $w_*^{(1)} = v_1^\top w_*$ is the minimum of $h(w^{(1)})$. Without loss of generality, assume

$$w_*^{(1)} > 0 = w_0^{(1)}.$$

Then by Lemma 3, we know the optimization path of Eq. (25) lies between $(0, w_*^{(1)})$, and for any $v \in (0, w_*^{(1)})$, we have

$$\alpha v - b \leq \nabla h(v) \leq \beta v - b, \quad b = -\nabla h(0).$$

Thus for Eq. (25) we have

$$\begin{aligned} w_{k+1}^{(1)} - w_k^{(1)} &= -\eta\nabla h(w_k^{(1)}) \leq -\eta(\alpha w_k^{(1)} - b), \\ w_{k+1}^{(1)} - w_k^{(1)} &= -\eta\nabla h(w_k^{(1)}) \geq -\eta(\beta w_k^{(1)} - b). \end{aligned}$$

Define the following dynamics:

$$u_{k+1}^{(1)} - u_k^{(1)} = -\eta(\alpha u_k^{(1)} - b), \quad v_{k+1}^{(1)} - v_k^{(1)} = -\eta(\beta v_k^{(1)} - b), \quad u_0^{(1)} = v_0^{(1)} = 0.$$

By the discrete Gronwall's inequality (Clark, 1987), we have

$$v_k^{(1)} \leq w_k^{(1)} \leq u_k^{(1)}.$$

Furthermore, $u_k^{(1)}$ and $v_k^{(1)}$ satisfy two first order recurrence relations respectively, thus they can be solved by

$$u_k^{(1)} = \eta \sum_{i=1}^{k} (1-\eta\alpha)^{i-1} b, \quad v_k^{(1)} = \eta \sum_{i=1}^{k} (1-\eta\beta)^{i-1} b.$$

Since $\eta < \frac{1}{\beta} \leq \frac{1}{\alpha}$, $u_k^{(1)}$ and $v_k^{(1)}$ converge. And $w_k^{(1)}$ also converges since $h(\cdot)$ is β-smooth convex and $\eta < \frac{1}{\beta}$.

In a same way, for the regularized path,

$$\hat{w}_{k+1,\lambda}^{(1)} = \hat{w}_{k,\lambda}^{(1)} - \gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}), \quad \hat{w}_{0,\lambda}^{(1)} = 0,$$

we have

$$\begin{aligned}
\hat{w}_{k+1,\lambda}^{(1)} - \hat{w}_{k,\lambda}^{(1)} &= -\gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}) \leq -\gamma\left((\alpha+\lambda)\hat{w}_{k,\lambda}^{(1)} - b\right), \\
\hat{w}_{k+1,\lambda}^{(1)} - \hat{w}_{k,\lambda}^{(1)} &= -\gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}) \geq -\gamma\left((\beta+\lambda)\hat{w}_{k,\lambda}^{(1)} - b\right).
\end{aligned}$$

Consider the following dynamics:

$$\hat{u}_{k+1,\lambda}^{(1)} - \hat{u}_{k,\lambda}^{(1)} = -\gamma((\alpha+\lambda)\hat{u}_{k,\lambda}^{(1)} - b), \quad \hat{v}_{k+1,\lambda}^{(1)} - \hat{v}_{k,\lambda}^{(1)} = -\gamma((\beta+\lambda)\hat{v}_{k,\lambda}^{(1)} - b),$$

where $\hat{u}_{0,\lambda}^{(1)} = \hat{v}_{0,\lambda}^{(1)} = 0$. Then by the discrete Gronwall's inequality (Clark, 1987) and the solution of the first order recurrence relation we obtain

$$\hat{v}_{k,\lambda}^{(1)} \leq \hat{w}_{k,\lambda}^{(1)} \leq \hat{u}_{k,\lambda}^{(1)}, \quad \hat{u}_{k,\lambda}^{(1)} = \gamma \sum_{i=1}^{k} (1-\gamma(\alpha+\lambda))^{i-1} b, \quad \hat{v}_{k,\lambda}^{(1)} = \gamma \sum_{i=1}^{k} (1-\gamma(\beta+\lambda))^{i-1} b.$$

Now we turn to bound the iterate averaged solution. Consider

$$\lambda_1 = \frac{1}{\gamma} - \frac{1}{\eta} + \beta - \alpha, \quad \lambda_2 = \frac{1}{\gamma} - \frac{1}{\eta} + \alpha - \beta,$$

since $\beta \geq \alpha$ and $0 < \gamma < \frac{1}{\beta - \alpha + 1/\eta}$ we know $\lambda_1 \geq \lambda_2 > 0$. Notice that

$$0 < \gamma(\alpha+\lambda_2) \leq \{\gamma(\alpha+\lambda_1),\ \gamma(\beta+\lambda_2)\} \leq \gamma(\beta+\lambda_1) = 1 - \gamma(-\frac{1}{\eta} + 2\beta - \alpha) < 1,$$

where the last inequality is because $\eta > \frac{1}{2\beta-\alpha}$. Thus $\hat{u}_{k,\lambda_1}^{(1)}$, $\hat{u}_{k,\lambda_2}^{(1)}$, $\hat{v}_{k,\lambda_1}^{(1)}$, $\hat{v}_{k,\lambda_2}^{(1)}$ converge. Further $\hat{w}_{k,\lambda_1}$ and $\hat{w}_{k,\lambda_2}$ also converge since $\gamma < \frac{1}{\beta+\lambda_1} \leq \frac{1}{\beta+\lambda_2}$ and the corresponding regularized losses are $(\beta+\lambda_1)$ and $(\beta+\lambda_2)$-smooth, respectively.

Next let us consider the weighting scheme $P_k = 1 - \left(\frac{\gamma}{\eta}\right)^{k+1}$, which is well defined since $0 < \gamma < \frac{1}{\beta-\alpha+1/\eta} \leq \eta$.

One can directly verify that $\tilde{u}_k^{(1)} = \frac{1}{P_k}\sum_{i=1}^{k} p_i u_i^{(1)}$, $\tilde{v}_k^{(1)} = \frac{1}{P_k}\sum_{i=1}^{k} p_i v_i^{(1)}$ converge, and

$$(1-P_k)(u_{k+1}^{(1)} - u_k^{(1)}) = \hat{v}_{k+1,\lambda_2}^{(1)} - \hat{v}_{k,\lambda_2}^{(1)}, \quad (1-P_k)(v_{k+1}^{(1)} - v_k^{(1)}) = \hat{u}_{k+1,\lambda_1}^{(1)} - \hat{u}_{k,\lambda_1}^{(1)}.$$

Thus according to Lemma 2 we have

$$P_k(u_k^{(1)} - \tilde{u}_k^{(1)}) = u_k^{(1)} - \hat{v}_{k,\lambda_2}^{(1)}, \quad P_k(v_k^{(1)} - \tilde{v}_k^{(1)}) = v_k^{(1)} - \hat{u}_{k,\lambda_1}^{(1)}.$$

Therefore

$$\begin{aligned}
\tilde{w}_k^{(1)} - \hat{w}_{k,\lambda_2}^{(1)} &\leq \tilde{u}_k^{(1)} - \hat{v}_{k,\lambda_2}^{(1)} = \tilde{u}_k^{(1)} - u_k^{(1)} + P_k(u_k^{(1)} - \tilde{u}_k^{(1)}) = (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}), \\
\tilde{w}_k^{(1)} - \hat{w}_{k,\lambda_1}^{(1)} &\geq \tilde{v}_k^{(1)} - \hat{u}_{k,\lambda_1}^{(1)} = \tilde{v}_k^{(1)} - v_k^{(1)} + P_k(v_k^{(1)} - \tilde{v}_k^{(1)}) = (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}),
\end{aligned}$$

which implies that

$$\hat{w}_{k,\lambda_1}^{(1)} + (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}) \leq \tilde{w}_k^{(1)} \leq \hat{w}_{k,\lambda_2}^{(1)} + (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}). \tag{26}$$

Note that $u_k^{(1)},\ \tilde{u}_k^{(1)},\ v_k^{(1)},\ \tilde{v}_k^{(1)},\ \hat{w}_{k,\lambda_1}^{(1)},\ \hat{w}_{k,\lambda_2}^{(1)}$ converge, therefore there is a constant M controlling their ℓ_2-norm. Define $m_k^{(1)} = (\hat{w}_{k,\lambda_2}^{(1)} + \hat{w}_{k,\lambda_1}^{(1)})/2$, $d_k^{(1)} = (\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{k,\lambda_1}^{(1)})/2$. Recall that $\hat{w}_{k,\lambda_1}^{(1)}$ are the GD optimization path of a $(\alpha+\lambda_1)$-strongly convex and $(\beta+\lambda_1)$-smooth loss, thus $\hat{w}_{k,\lambda_1}^{(1)}$ converges in rate $\mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right)$. Similarly $\hat{w}_{k,\lambda_2}^{(1)}$ converges in rate $\mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right)$. Thus triangle inequality we have

$$\left\|m_k^{(1)} - m^{(1)}\right\|_2 \le \frac{1}{2}\left\|\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{\infty,\lambda_2}^{(1)}\right\|_2 + \frac{1}{2}\left\|\hat{w}_{k,\lambda_1}^{(1)} - \hat{w}_{\infty,\lambda_1}^{(1)}\right\|_2 \le \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right).$$

$$\left\|d_k^{(1)} - d^{(1)}\right\|_2 \le \frac{1}{2}\left\|\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{\infty,\lambda_2}^{(1)}\right\|_2 + \frac{1}{2}\left\|\hat{w}_{k,\lambda_1}^{(1)} - \hat{w}_{\infty,\lambda_1}^{(1)}\right\|_2 \le \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right).$$

By Eq. (26) we obtain

$$\begin{aligned}
\tilde{w}_k^{(1)} - m_k^{(1)} &\le d_k^{(1)} + (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}) \le d_k^{(1)} + 2M\left(\frac{\gamma}{\eta}\right)^{k+1} \\
&\le d^{(1)} - d^{(1)} + d_k^{(1)} + \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right) \\
&\le d^{(1)} + \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right) + \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right),
\end{aligned}$$

and

$$\begin{aligned}
\tilde{w}_k^{(1)} - m_k^{(1)} &\ge d_k^{(1)} + (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}) \ge d_k^{(1)} - 2M\left(\frac{\gamma}{\eta}\right)^{k+1} \\
&\ge d^{(1)} - d^{(1)} + d_k^{(1)} - \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right) \\
&\ge d^{(1)} - \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) - \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right) - \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right).
\end{aligned}$$

Thus

$$\left\|\tilde{w}_k^{(1)} - m_k^{(1)}\right\|_2 \le d^{(1)} + \mathcal{O}(C^k), \quad C = \max\{(1-\gamma(\alpha+\lambda_1), (1-\gamma(\alpha+\lambda_2), \frac{\gamma}{\eta}\}.$$

In conclusion we have

$$\left\|\tilde{w}_k^{(1)} - m^{(1)}\right\|_2 \le \left\|\tilde{w}_k^{(1)} - m_k^{(1)}\right\|_2 + \left\|m_k^{(1)} - m^{(1)}\right\|_2 \le d^{(1)} + \mathcal{O}(C^k).$$

□

D. Experiments setups

The code is available at `https://github.com/uuujf/IterAvg`.

The experiments are conducted using one GPU K80 and `PyTorch 1.3.1`.

D.1. Two dimensional toy example

The loss function is

$$\begin{aligned}
&L(w) = \frac{1}{2}(w-w_*)^\top \Sigma (w-w_*), \quad w_* = (1,1)^\top, \quad \Sigma = U\,\mathrm{Diag}\,(0.1, 1)\,U^T, \\
&U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{\pi}{3}.
\end{aligned}$$

All the algorithms are initiated from zero. The learning rate for the unregularized problem is $\eta = 0.1$. The hyperparameter for the vanilla/generalized ℓ_2-regularization is $\lambda = 0.1$. And the learning rate for the regularized problem is $\gamma = \frac{1}{\lambda + 1/\eta}$. The preconditioning matrix is set to be $Q = \Sigma$. We run the algorithms for 500 iterations. For NGD and NSGD, we set the strongly convex coefficient to be $\alpha = 0.05$.

D.2. MNIST dataset

Dataset http://yann.lecun.com/exdb/mnist/

Linear regression The image data is scaled to $[0, 1]$. The label data is one-hotted. The loss function is standard linear regression under squared loss, without bias term, $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^T x_i - y_i\right\|_2^2$. All the algorithms are initiated from zero. The learning rate for the unregularized problem is $\eta = 0.01$. The hyperparameter for the vanilla/generalized ℓ_2-regularizer is $\lambda = 4.0$. And the learning rate for the regularized problem is $\gamma = \frac{1}{\lambda + 1/\eta}$. The preconditioning matrix is set to be $Q = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$. The batch size for the stochastic algorithms are $b = 500$. We run the algorithms for 500 iterations. For NGD and NSGD, we set the strongly convex coefficient to be $\alpha = 1.0$.

Logistic regression The image data is scaled to $[0, 1]$. The label data is one-hotted. The loss function is standard logistics regression loss plus an ℓ_2-regularization term, $L(w) = \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}(y_i \mid\mid \sigma(w^\top x_i)) + \frac{\lambda_0}{2}\|w\|_2^2$, where $\sigma(x)$ is the softmax function and $\lambda_0 = 1.0$. All the algorithms are initiated from zero. The learning rate for the unregularized problem is $\eta = 0.01$. The hyperparameter for the vanilla/generalized ℓ_2-regularizer is $\lambda = 4.0$. And the learning rate for the regularized problem is $\gamma = \frac{1}{\lambda + 1/\eta}$. The preconditioning matrix is set to be $Q = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$. The batch size for the stochastic algorithms are $b = 500$. We run the algorithms for 500 iterations. For NGD and NSGD, we set the strongly convex coefficient to be $\alpha = 1.0$.

D.3. CIFAR-10 and CIFARR-100 datasets

Datasets https://www.cs.toronto.edu/~kriz/cifar.html

VGG-16 on CIFAR-10 The image data is scaled to $[0, 1]$ and augmented by horizontally flipping and randomly cropping. The label data is one-hotted. The model is standard VGG-16 with batch normalization. We train the model with vanilla SGD for 300 epochs. The batch size is 100. The learning rate is 0.1, and decreased by ten times at epoch 150 and 250. The weight decay is set to be 5×10^{-4}.

After finishing the SGD training process, we average the checkpoints from 61 to 300 epoch with standard geometric distribution. We test the success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$. And the best one is 0.99.

ResNet-18 on CIFAR-10 The image data is scaled to $[0, 1]$ and augmented by horizontally flipping and randomly cropping. The label data is one-hotted. The model is standard ResNet-18. We train the model with vanilla SGD for 300 epochs. The batch size is 100. The learning rate is 0.1, and decreased by ten times at epoch 150 and 250. The weight decay is set to be 5×10^{-4}.

After finishing the SGD training process, we average the checkpoints from 61 to 300 epoch with standard geometric distribution. We test the success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$. And the best one is 0.99.

ResNet-18 on CIFAR-100 The image data is scaled to $[0, 1]$ and augmented by horizontally flipping and randomly cropping. The label data is one-hotted. The model is standard ResNet-18. We train the model with vanilla SGD for 300 epochs. The batch size is 100. The learning rate is 0.1, and decreased by ten times at epoch 150 and 250. The weight decay is set to be 5×10^{-4}.

After finishing the SGD training process, we average the checkpoints from 61 to 300 epoch with standard geometric distribution. We test the success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$. And the best one is 0.99.

Additional experiments for deep nets without weight decay For ResNet-18 trained on CIFAR-10, without weight decay, and with the other setups the same, vanilla SGD has 92.95% test accuracy, and our method has 93.21% test accuracy. This result is consistent with the results presented in the main text.