A. Missing proofs in main paper

A.1. Proof of Proposition 1

Proof. We first calculate the expectation and variance of the sampling random vector $W_{\text{sgd}}$, then obtain that of the sampling noise $V_{\text{sgd}}$.

**Sampling with replacement** In the circumstance of sampling with replacement, the sampling random vector $W_{\text{sgd}}$ could be decompose as

$$W_{\text{sgd}} = W^1 + \cdots + W^b,$$

where $W^1, \ldots, W^b$ are i.i.d. and each of them represents once sampling procedure. Thus $W_i = (w_i^1, \ldots, w_i^n)^T$ contains one multiple of $\frac{1}{n}$ and $n-1$ multiples of zero, with random index. Hence we have

$$\mathbb{E}[w_i^j] = \frac{1}{bn}, \quad \mathbb{E}[w_i^j w_i^k] = \frac{1}{b^2 n^2}, \quad \mathbb{E}[w_i^j w_i^k] = 0, \forall j \neq k.$$

Thus

$$\mathbb{E}[W_i] = \frac{1}{bn} \mathbb{I},$$

$$\text{Var}[W_i] = \mathbb{E}[W_i (W_i)^T] - \mathbb{E}[W_i] \mathbb{E}[W_i]^T = \left( \frac{1}{b^2 n} \right) - \frac{1}{b^2 n^2} \mathbb{I}^T = \frac{1}{b^2 n} \left( I - \frac{1}{n} \mathbb{I}^T \right).$$

Recall $W^1, \ldots, W^b$ are i.i.d., thus

$$\mathbb{E}[W_{\text{sgd}}] = b \mathbb{E}[W_i] = \frac{1}{n} \mathbb{I}, \quad \text{Var}[W_{\text{sgd}}] = b \text{Var}[W_i] = \frac{1}{bn} \left( I - \frac{1}{n} \mathbb{I}^T \right).$$

Therefore for the sampling noise $V_{\text{sgd}} = W_{\text{sgd}} - \frac{1}{n} \mathbb{I}$ we have

$$\mathbb{E}[V_{\text{sgd}}] = 0, \quad \text{Var}[V_{\text{sgd}}] = \frac{1}{bn} \left( I - \frac{1}{n} \mathbb{I}^T \right).$$

**Sampling without replacement** Let $W_{\text{sgd}}' = (w_1', \ldots, w_n')^T$. In the case of sampling without replacement, we know the sampling random vector $W_{\text{sgd}}'$ contains exactly $b$ multiples of $\frac{1}{n}$’s and $n-b$ multiples of zero, with random index. Hence we have

$$\mathbb{E}[w_i'] = \left( \frac{n-1}{b-1} \right) \left( \frac{1}{n} \right) = \frac{1}{n}, \quad \mathbb{E}[(w_i')^2] = \left( \frac{n-1}{b-1} \right) \left( \frac{1}{n} \right) = \frac{1}{bn}, \quad \mathbb{E}[w_i' w_i''] = \left( \frac{n-2}{b} \right) \left( \frac{1}{n} \right) = \frac{b-1}{bn(n-1)}, \forall j \neq k.$$

Thus

$$\mathbb{E}[W_{\text{sgd}}'] = \frac{1}{n} \mathbb{I},$$

$$\text{Var}[W_{\text{sgd}}'] = \mathbb{E}[W_{\text{sgd}}' (W_{\text{sgd}}')^T] - \mathbb{E}[W_{\text{sgd}}'] \mathbb{E}[W_{\text{sgd}}']^T$$

$$= \begin{pmatrix}
\frac{1}{bn} & \frac{1}{bn(n-1)} & \cdots & \frac{1}{bn(n-1)} \\
\frac{1}{bn(n-1)} & \frac{1}{bn} & \cdots & \frac{1}{bn(n-1)} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{bn(n-1)} & \frac{1}{bn(n-1)} & \cdots & \frac{1}{bn}
\end{pmatrix} - \frac{1}{n^2} \mathbb{I}^T = \frac{n-b}{bn(n-1)} \left( I - \frac{1}{n} \mathbb{I}^T \right).$$

Therefore for the sampling noise $V_{\text{sgd}} = W_{\text{sgd}} - \frac{1}{n} \mathbb{I}$ we have

$$\mathbb{E}[V_{\text{sgd}}'] = 0, \quad \text{Var}[V_{\text{sgd}}'] = \frac{n-b}{bn(n-1)} \left( I - \frac{1}{n} \mathbb{I}^T \right).$$
A.2. Proof of Theorem 1

Proof. Let

\[ \epsilon_n = y_n - x^T \theta, \]

by assumption we have

\[ \mathbb{E}[\epsilon_n x_n] = 0, \quad \mathbb{E}[\epsilon_n] = 0, \quad \mathbb{E}[\epsilon^2 x x^T] \leq \sigma^2 \Sigma. \]

Recall the MSGD updates

\[ \theta_{n+1} = \theta_n - \eta \sum_{r \in B} w_r (x_r x_r^T \theta_n - y_r x_r), \]

hence we have

\[ \theta_{n+1} - \theta_* = \left( I - \eta \sum_{r \in B} w_r x_r x_r^T \right) (\theta_n - \theta_*) + \eta \sum_{r \in B} w_r \epsilon_r x_r. \]

Define

\[ L(k) = \sum_{r \in B_k} w_r x_r x_r^T, \quad \text{(5)} \]

\[ M(i, k) = \left\{ \begin{array}{ll} (I - \eta L(i)) \cdots (I - \eta L(k)), & i \geq k \\ I, & i < k. \end{array} \right. \quad \text{(6)} \]

\[ N(k) = \sum_{r \in B_k} w_r \epsilon_r x_r, \quad \text{(7)} \]

Then recursively we obtain

\[ \theta_i - \theta_* = M(i, 1)(\theta_0 - \theta_*) + \eta \sum_{k=1}^i M(i, k+1) N(k). \quad \text{(8)} \]

Moments of \( L(k) \) We first calculate the first and second moments of \( N(k) \) defined in Eq. (7). Since \( \mathbb{E}[w_r] = 1/B, \mathbb{E}[w_r^2] = 1/bB, \mathbb{E}[w_i w_j] = \frac{b-1}{bB(B-1)} \), \( i \neq j \), and \( \mathbb{E} \left[ \|x\|^2_2 x x^T \right] \leq R^2 \Sigma, \Sigma \preceq \Lambda I \), we have

\[ \mathbb{E} \left[ L(k) \right] = \sum_{r=1}^B \mathbb{E}[w_r] \cdot \mathbb{E}[x_r x_r^T] = B \cdot \frac{1}{B} \cdot \Sigma = \Sigma. \]

\[ \mathbb{E} \left[ L(k)^2 \right] = B \sum_{r=1}^B \mathbb{E}[w_r^2] \cdot \mathbb{E}[x_r x_r^T x_r x_r^T] + 2B \sum_{r=1}^{B-1} \mathbb{E}[w_r w_s] \cdot \mathbb{E}[x_r x_r^T x_s x_s^T]
\]

\[ = B \sum_{r=1}^B \mathbb{E}[w_r] \cdot \mathbb{E} \left[ \|x_r\|^2_2 x_r x_r^T \right] + 2B \sum_{r=1}^{B-1} \mathbb{E}[w_r w_s] \cdot \mathbb{E}[x_r x_r^T] \cdot \mathbb{E}[x_s x_s^T]
\]

\[ = B \cdot \frac{1}{bB} \cdot \mathbb{E} \left[ \|x\|^2_2 x x^T \right] + 2B(B-1) \cdot \frac{b-1}{bB(B-1)} \cdot \Sigma^2
\]

\[ \leq \frac{1}{b} R^2 \Sigma + \frac{b-1}{b} \lambda \Sigma = \frac{R^2 + (b-1)\lambda}{b} \Sigma. \]

Moments of \( M(i, k) \) We only consider \( i \geq k \).

\[ \mathbb{E}[M(i, k)] = (I - \eta \mathbb{E}[L(i)]) \cdots (I - \eta \mathbb{E}[L(k)]) = (I - \eta \Sigma)^{i-k+1}. \]

\[ \mathbb{E}M(i, k)M(i, k)^T = \mathbb{E}(I - \eta L(i))^2 (I - \eta L(i))^2 M(i, k+1)^T
\]

\[ \leq \mathbb{E}M(i, k+1) \left( I - 2\eta L(k) + \eta^2 L(k)^2 \right) M(i, k+1)^T
\]

\[ = \mathbb{E}M(i, k+1)(I - 2\eta \Sigma + \eta^2 \frac{R^2 + (b-1)\lambda}{b} \Sigma) M(i, k+1)^T
\]

\[ = \mathbb{E}M(i, k+1)M(i, k+1)^T - \eta \left( 2 - \eta \frac{R^2 + (b-1)\lambda}{b} \right) \mathbb{E}M(i, k+1)M(i, k+1)^T. \]
Hence
\[
E M(i, k + 1) \Sigma M(i, k + 1)^T \leq \frac{1}{\eta \left(2 - \eta R^2 + (b-1)\lambda \right)} \left( E M(i, k + 1) M(i, k + 1)^T - E M(i, k) M(i, k)^T \right).
\]

**Moments of \(N(k)\)**

\[
E[N(k)] = \sum_{r \in B_k} E[w_r] \cdot E[\epsilon_r] = B \cdot \frac{1}{B} \cdot 0 = 0.
\]

\[
E \left[ N(k) N(k)^T \right] = E \sum_{r=1}^{B} w_r^2 \epsilon_r^2 x_r^2 x_r^T + 2E \sum_{r=1}^{B-1} \sum_{s=2}^{B} w_r w_s \epsilon_r \epsilon_s x_r x_s^T
\]

\[
= \sum_{r=1}^{B} E[w_r^2] \cdot E[\epsilon_r^2] x_r^2 x_r^T + 2 \sum_{r=1}^{B-1} \sum_{s=2}^{B} E[w_r] \cdot E[w_s] \cdot E[\epsilon_r] \cdot E[\epsilon_s] x_r x_s^T
\]

\[
= B \cdot \frac{1}{bB} \cdot E[\epsilon^2] x_r x_r^T + 2 \frac{B(B-1)}{2} \cdot \frac{1}{B^2} \cdot 0
\]

\[
\leq \frac{1}{b} \sigma^2 \Sigma.
\]

**Calculate averaging**

Taking expectation to \(w_k\) and \(B_k\), we have

\[
E \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \langle \theta_i - \theta_*, \Sigma(\theta_j - \theta_*) \rangle
\]

\[
= E \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \left( \theta_i - \theta_*, \Sigma \left( M(j, i + 1)(\theta_i - \theta_*) + \eta \sum_{k=i+1}^{j} M(j, k + 1)N(k) \right) \right)
\]

\[
= E \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \langle \theta_i - \theta_*, \Sigma M(j, i + 1)(\theta_i - \theta_*) \rangle
\]

\[
= E \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \langle \theta_i - \theta_*, \Sigma (I - \eta \Sigma) \rangle \langle \theta_i - \theta_*, \eta^{-1} \rangle
\]

\[
= E \sum_{i=0}^{n} \langle \theta_i - \theta_*, \eta^{-1} (I - \eta \Sigma) \rangle \langle \theta_i - \theta_*, \eta^{-1} \rangle
\]

\[
= \eta^{-1} E \sum_{i=0}^{n} \| \theta_i - \theta_* \|_2^2 - E \sum_{i=0}^{n} \| \Sigma^{1/2} (\theta_i - \theta_*) \|_2^2,
\]

which implies that

\[
(n + 1)^2 E \| \Sigma^{1/2} (\theta_n - \theta_*) \|_2^2 = E \sum_{i,j=0}^{n} \langle \theta_i - \theta_*, \Sigma(\theta_j - \theta_*) \rangle
\]

\[
= E \sum_{i=0}^{n} \langle \theta_i - \theta_*, \Sigma(\theta_i - \theta_*) \rangle + 2E \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \langle \theta_i - \theta_*, \Sigma(\theta_j - \theta_*) \rangle
\]

\[
\leq E \sum_{i=0}^{n} \| \Sigma^{1/2} (\theta_i - \theta_*) \|_2^2 + 2\eta^{-1} E \sum_{i=0}^{n} \| \theta_i - \theta_* \|_2^2 - 2E \sum_{i=0}^{n} \| \Sigma^{1/2} (\theta_i - \theta_*) \|_2^2
\]

\[
\leq 2\eta^{-1} E \sum_{i=0}^{n} \| \theta_i - \theta_* \|_2^2,
\]
and in the following we bound $E \sum_{i=0}^{n} \|\theta_i - \theta_*\|_2^2$. We do so by bounding each term.

Now since the solution of $\theta_i$ in Eq. (8) and the fact $E[N(k)] = 0$, we have

$$E \|\theta_i - \theta_*\|_2^2 = E \|M(i, 1)(\theta_0 - \theta_*)\|_2^2 + \eta^2 E \sum_{k=1}^{i} \sum_{j=1}^{i} \langle M(i, k + 1)N(k), M(i, j + 1)N(j)\rangle$$

$$= E \|M(i, 1)(\theta_0 - \theta_*)\|_2^2 + \eta^2 E \sum_{k=1}^{i} \langle M(i, k + 1)N(k), M(i, k + 1)N(k)\rangle.$$ 

In conclusion we have

$$\frac{1}{2} \eta(n + 1)^2 E \left\| \sum_{n=0}^{\infty} (\tilde{\theta}_n - \theta_*) \right\|_2^2 \leq E \sum_{i=0}^{n} \|\theta_i - \theta_*\|_2^2$$

$$= E \sum_{i=0}^{n} \|M(i, 1)(\theta_0 - \theta_*)\|_2^2 + \eta^2 E \sum_{i=0}^{n} \sum_{k=1}^{i} \langle M(i, k + 1)N(k), M(i, k + 1)N(k)\rangle.$$ 

We call the two terms as the noiseless term and the noise term.

**Noise term** We bound the noise term by observing that

$$E \langle M(i, k + 1)N(k), M(i, k + 1)N(k)\rangle = E M(i, k + 1)N(k)N(k)^T M(i, k + 1)^T$$

$$= E \text{Tr} \left[ N(k)N(k)^T M(i, k + 1)^2 M(i, k + 1) \right] = \text{Tr} \left[ E \left[ N(k)N(k)^T M(i, k + 1) \right] \right]$$

$$\leq \sigma^2 \text{Tr} \left[ \Sigma \cdot E \left[ M(i, k + 1)^T M(i, k + 1) \right] \right] = \sigma^2 \text{Tr} \left[ M(i, k + 1)\Sigma M(i, k + 1)^T \right]$$

$$\leq \frac{\sigma^2}{\eta \left(2 - \eta \frac{R^2 + (b-1)\lambda}{b}\right)} \text{Tr} \left( EM(i, k + 1)M(i, k + 1)^T - EM(i, k)M(i, k)^T \right).$$

Hence

$$\eta^2 E \sum_{i=0}^{n} \sum_{k=1}^{i} \langle M(i, k + 1)N(k), M(i, k + 1)N(k)\rangle$$

$$\leq \frac{\eta \sigma^2}{\left(2 - \eta \frac{R^2 + (b-1)\lambda}{b}\right)} \sum_{i=0}^{n} \sum_{k=1}^{i} \text{Tr} \left( EM(i, k + 1)M(i, k + 1)^T - EM(i, k)M(i, k)^T \right)$$

$$= \frac{\eta \sigma^2}{\left(2 - \eta \frac{R^2 + (b-1)\lambda}{b}\right)} \sum_{i=0}^{n} \text{Tr} \left( EM(i, i + 1)M(i, i + 1)^T - EM(i, i)M(i, i)^T \right)$$

$$\leq \frac{\eta \sigma^2}{\left(2 - \eta \frac{R^2 + (b-1)\lambda}{b}\right)} \sum_{i=0}^{n} \text{Tr} \left( EM(i, i + 1)M(i, i + 1)^T \right)$$

$$= \frac{\eta \sigma^2}{\left(2 - \eta \frac{R^2 + (b-1)\lambda}{b}\right)} (n + 1) \text{Tr}[J] = \frac{\eta \sigma^2}{\left(2 - \eta \frac{R^2 + (b-1)\lambda}{b}\right)} (n + 1)d.$$

**Noiseless term** Let $E_0 = (\theta_0 - \theta_*)(\theta_0 - \theta_*)^T$. Define two linear operators $S$ and $T$ from symmetric matrices to symmetric matrices as

$$SA = E[L(k)AL(k)]$$

$$TA = \Sigma A + A\Sigma - \eta E[L(k)AL(k)] = \Sigma A + A\Sigma - \eta SA.$$
With these notations and $M(i, 1) = (I - \eta L(i)) \cdots (I - \eta L(1))$, we recursively have
\[ E[M(i, 1)^T M(i, 1)] = (I - \eta T)^i I. \]

Next we bound the noiseless term
\[ E \sum_{i=0}^{n} \|M(i, 1)(\theta_0 - \theta_*)\|_2^2 = E \sum_{i=0}^{n} \text{Tr}[M(i, 1)^T M(i, 1)(\theta_0 - \theta_*)(\theta_0 - \theta_*)^T] \]
\[ = \sum_{i=0}^{n} \langle EM(i, 1)^T M(i, 1), E_0 \rangle \]
\[ = \sum_{i=0}^{n} \langle (I - \eta T)^i I, E_0 \rangle \]
\[ = \langle \eta^{-1} T^{-1} (I - (I - \eta T)^{n+1}) I, E_0 \rangle \]
\[ \leq \langle \eta^{-1} T^{-1} I, E_0 \rangle. \]

Let $M = T^{-1} I$, then $I = TM = \Sigma M + M \Sigma - \eta SM$, hence by the Kronecker’s produce we have
\[ I + \eta SM = \Sigma M + M \Sigma = (\Sigma \otimes I + I \otimes \Sigma) M, \]
thus
\[ M = (\Sigma \otimes I + I \otimes \Sigma)^{-1} I + (\Sigma \otimes I + I \otimes \Sigma)^{-1} \eta SM = \frac{1}{2} \Sigma^{-1} + (\Sigma \otimes I + I \otimes \Sigma)^{-1} \eta SM. \]

Therefore
\[ E \sum_{i=0}^{n} \|M(i, 1)(\theta_0 - \theta_*)\|_2^2 = \langle \eta^{-1} M, E_0 \rangle = \frac{1}{2\eta} \langle \Sigma^{-1}, E_0 \rangle + \langle (\Sigma \otimes I + I \otimes \Sigma)^{-1} SM, E_0 \rangle \]
\[ = \frac{1}{2\eta} \langle \theta_0 - \theta_* \rangle^T \Sigma^{-1} (\theta_0 - \theta_*) + \langle SM, (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \rangle. \]

We left to bound $SM$ and $(\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0$.

**Bound $(\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0$** By Cauchy-Schwarz inequality we have
\[ E_0 = \Sigma^\frac{1}{2} \Sigma^{-\frac{1}{2}} (\theta_0 - \theta_*) (\theta_0 - \theta_*)^T \Sigma^{-\frac{1}{2}} \Sigma^\frac{1}{2} \leq (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \Sigma. \]
Thus
\[ (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \leq (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \Sigma = (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \frac{1}{2} I. \]

**Bound $SM$** Firstly by definition,
\[ \text{Tr}[SM] = E \text{Tr}[L(k)ML(k)] = E \sum_{r=1}^{B} \text{Tr}[w_r^2 x_r x_r^T M x_r x_r^T] + 2E \sum_{r=1}^{B-1} \sum_{s=2}^{B} \text{Tr}[w_r w_s x_r x_r^T M x_s x_s^T] \]
\[ = \sum_{r=1}^{B} E[w_r^2] \cdot \text{Tr} \left[ \left\| x_r \right\|_2^2 x_r x_r^T \right] M \right] + 2 \sum_{r=1}^{B-1} \sum_{s=2}^{B} E[w_r w_s] \cdot \text{Tr}[E[x_r x_r^T] \cdot M \cdot E[x_s x_s^T]] \]
\[ \leq B \cdot \frac{1}{b^2} \cdot \text{Tr} \left[ R^2 \Sigma M \right] + 2 \frac{B(B - 1)}{2} \cdot \frac{b - 1}{bB(B - 1)} \cdot \text{Tr}[\Sigma M \Sigma] \]
\[ \leq \frac{R^2}{b} \cdot \text{Tr}[\Sigma M] + \frac{b - 1}{b} \cdot \lambda \cdot \text{Tr}[M \Sigma] = \frac{R^2 + (b - 1)\lambda}{b} \text{Tr}[\Sigma M]. \]
Secondly, taking trace we have:

\[ d = \text{Tr}[I] = \text{Tr}[TM] = 2 \text{Tr}[\Sigma M] - \eta \text{Tr}[SM] \geq 2 \text{Tr}[\Sigma M] \geq \frac{2b}{R^2 + (b - 1)\lambda} \text{Tr}[SM], \]

which implies that \( \text{Tr}[SM] \leq \frac{R^2 + (b - 1)\lambda}{2b} d. \)

To sum up we have:

\[
\langle SM, (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \rangle \\
\leq \frac{1}{2} (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \langle SM, I \rangle \\
= \frac{1}{2} (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \text{Tr}[SM] \\
\leq \frac{(R^2 + (b - 1)\lambda)d}{4b} (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*). 
\]

Therefore for the noiseless term we have:

\[
\mathbb{E} \sum_{i=0}^{n} \| M(i, 1)(\theta_0 - \theta_*) \|^2 \leq \frac{1}{2\eta} (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) + \langle SM, (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \rangle \\
\leq \left( \frac{1}{2\eta} + \frac{(R^2 + (b - 1)\lambda)d}{4b} \right) (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*). 
\]

In conclusion we have:

\[
\frac{1}{2} \eta(n + 1)^2 \mathbb{E} \left\| \Sigma^{\frac{1}{2}} (\tilde{\theta}_n - \theta_*) \right\|^2 \leq \text{noiseless term + noise term} \\
\leq \frac{\eta \sigma^2}{\left( 2 - \eta \frac{R^2 + (b - 1)\lambda}{b} \right)} (n + 1)d + \left( \frac{1}{2\eta} + \frac{(R^2 + (b - 1)\lambda)d}{4b} \right) (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*). 
\]

Hence:

\[
\mathbb{E} \left\| \Sigma^{\frac{1}{2}} (\tilde{\theta}_n - \theta_*) \right\|^2 \leq \frac{1}{n + 1} \cdot \frac{2\sigma^2 d}{\left( 2 - \eta \frac{R^2 + (b - 1)\lambda}{b} \right)} + \frac{1}{(n + 1)^2} \left( 1 + \frac{(R^2 + (b - 1)\lambda)d}{2b} \right) (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*), 
\]

which complete our proof.

\[ \square \]

**B. Strong convergence of Gaussian MSGD and its SDE**

**Theorem 2.** (Strong convergence between Gaussian MSGD and SDE) Let \( T \geq 0. \) Let \( C(\theta) \) be the diffusion matrix, e.g., \( C(\theta) = \frac{1}{\sqrt{bN}} \nabla \theta \mathcal{L}(\theta) \in \mathbb{R}^{D \times N}. \) Assume there exist some \( L, M > 0 \) such that \( \max_{i=1,2,\ldots,N} \| \nabla \ell_i(\theta) \| \leq M \) and that \( \nabla \ell_i(\theta) \) are Lipschitz continuous with bounded Lipschitz constant \( L > 0 \) uniformly for all \( i = 1, 2, \ldots, N. \)

Then the Gaussian MSGD iteration (9)

\[ \theta_{k+1} - \theta_k = -\eta \nabla \theta L(\theta_k) + \eta C(\theta_k)W_{k+1}, \ W_k \sim \mathcal{N}(0, I), \ i.i.d. \]  

is a order 1 strong approximation to SDE (10)

\[ d\Theta_t = -\nabla \theta L(\Theta_t)dt + \sqrt{\eta} C(\Theta_t) dW_t, \ \Theta_0 = \theta_0, \ W_t \in \mathbb{R}^N \text{ is a standard Brownian motion} \]

i.e., there exist a constant \( C \) independent on \( \eta \) but depending on \( L \) and \( M \) such that

\[ \mathbb{E} \| \Theta_{k\eta} - \theta_k \|^2 \leq C \eta^2, \quad \text{for all } 0 \leq k \leq \lfloor T/\eta \rfloor. \]  

(11)
On the Noisy Gradient Descent that Generalizes as SGD

Proof. We show that, as \( \eta \to 0 \), the discrete iteration \( \theta_t \) of Eq. (9) in strong norm and on finite-time intervals is close to the solution of the SDE (10). The main techniques follow (Borkar & Mitter, 1999), but (Borkar & Mitter, 1999) only considered the case when \( C(\theta) \) is a constant.

For vector \( x \in \mathbb{R}^d \), we define its norm as \( |x| := \sqrt{x^T x} \); for matrix \( X \in \mathbb{R}^{d_1 \times d_2} \), we define its norm as \( |X| := \sqrt{\text{Tr}(X^TX)} = \sqrt{\text{Tr}(XX^T)} \).

Let \( \hat{\Theta}_t \) be the process defined by the integral form of the stochastic differential equation

\[
\hat{\Theta}_t - \hat{\Theta}_0 = -\int_0^t \nabla_\theta L(\hat{\Theta}_s, \eta_s) ds + \sqrt{\eta} \int_0^t C(\hat{\Theta}_s, \eta_s) dW_s, \quad \hat{\Theta}_0 = \theta_0.
\]  

(12)

Here for a real positive number \( a > 0 \) we define \( |a| = \max\{k \in \mathbb{N}_+, k < a\} \). From (12) we see that we have, for \( k = 0, 1, 2, ... \)

\[
\hat{\Theta}_{(k+1)\eta} - \hat{\Theta}_{k\eta} = -\eta \nabla_\theta L(\hat{\Theta}_{k\eta}) - \sqrt{\eta} C(\hat{\Theta}_{k\eta})(W_{(k+1)\eta} - W_{k\eta}).
\]  

(13)

Since \( \sqrt{\eta}(W_{(k+1)\eta} - W_{k\eta}) \sim \mathcal{N}(0, \eta^2 I) \), we could let \( \eta W_{k+1} = \sqrt{\eta}(W_{(k+1)\eta} - W_{k\eta}) \), where \( W_{k+1} \) is the i.i.d. Gaussian sequence in (9). From here, we see that

\[
\hat{\Theta}_{k\eta} = \theta_k,
\]  

(14)

where \( \theta_k \) is the solution to (9).

We first bound \( \hat{\Theta}_t \) in Eq. (12) and \( \Theta_t \) in Eq. (10). Then we could obtain the error estimation of \( \theta_k = \hat{\Theta}_{k\eta} \) and \( \Theta_{k\eta} \) by simply set \( t = k\eta \).

Since we assumed that \( \nabla_\theta \ell_i(\theta) \) is \( L \)-Lipschitz continuous, we get \( |C(\theta_1) - C(\theta_2)| = \frac{1}{\sqrt{bN}} \left( \sum_{i=1}^N |\nabla_\theta \ell_i(\theta_1) - \nabla_\theta \ell_i(\theta_2)|^2 \right) \leq \frac{1}{\sqrt{bN}} \sqrt{NL^2|\theta_1 - \theta_2|^2} \leq L|\theta_1 - \theta_2| \) since \( b \geq 1 \). Thus \( C(\theta) \) is also \( L \)-Lipschitz continuous. Take a difference between (12) and (10) we get

\[
\hat{\Theta}_t - \Theta_t = -\int_0^t [\nabla_\theta L(\Theta_s, \eta_s) - \nabla_\theta L(\Theta_s)] ds + \sqrt{\eta} \int_0^t [C(\hat{\Theta}_s, \eta_s) - C(\Theta_s)] dW_s.
\]  

(15)

We can estimate

\[
|\nabla_\theta L(\hat{\Theta}_{(k+1)\eta}) - \nabla_\theta L(\Theta_{k\eta})|^2 \leq 2|\nabla_\theta L(\hat{\Theta}_{(k+1)\eta}) - \nabla_\theta L(\Theta_{(k+1)\eta})|^2 + 2|\nabla_\theta L(\Theta_{(k+1)\eta}) - \nabla_\theta L(\Theta_{k\eta})|^2 \leq 2L^2|\Theta_{(k+1)\eta} - \Theta_{(k+1)\eta}|^2 + 2L^2|\Theta_{(k+1)\eta} - \Theta_{k\eta}|^2,
\]  

(16)

where we used the inequality \( |\nabla_\theta L(\theta_1) - \nabla_\theta L(\theta_2)| \leq \frac{1}{N} \sum_{i=1}^N |\nabla_\theta \ell_i(\theta_1) - \nabla_\theta \ell_i(\theta_2)| \leq L|\theta_1 - \theta_2| \).

Similarly, we estimate

\[
|C(\hat{\Theta}_{(k+1)\eta}) - C(\Theta_{k\eta})|^2 \leq 2|C(\hat{\Theta}_{(k+1)\eta}) - C(\Theta_{(k+1)\eta})|^2 + 2|C(\Theta_{(k+1)\eta}) - C(\Theta_{k\eta})|^2 \leq 2L^2|\Theta_{(k+1)\eta} - \Theta_{(k+1)\eta}|^2 + 2L^2|\Theta_{(k+1)\eta} - \Theta_{k\eta}|^2.
\]  

(17)
On the Noisy Gradient Descent that Generalizes as SGD

On the other hand, from (15), the Itô’s isometry (Øksendal, 2003) and Cauchy–Schwarz inequality we have

\[
\mathbb{E}|\hat{\Theta}_t - \Theta_t|^2 \\
\leq 2\mathbb{E} \left[ \int_0^t |\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta} \rfloor}) - \nabla_\theta L(\Theta_s)|^2 ds \right] + 2\eta \mathbb{E} \left[ \int_0^t |C(\Theta_{\lfloor \frac{s}{\eta} \rfloor}) - C(\Theta_s)|^2 dW_s \right] \\
\leq 2\mathbb{E} \left[ \int_0^t |\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta} \rfloor}) - \nabla_\theta L(\Theta_s)|^2 ds \right] + 2\eta \int_0^t \mathbb{E} \left| C(\Theta_{\lfloor \frac{s}{\eta} \rfloor}) - C(\Theta_s) \right|^2 ds \\
\leq 2\int_0^t \mathbb{E} \left| \nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta} \rfloor}) - \nabla_\theta L(\Theta_s) \right|^2 ds + 2\eta \int_0^t \mathbb{E} \left| C(\Theta_{\lfloor \frac{s}{\eta} \rfloor}) - C(\Theta_s) \right|^2 ds. 
\]

Combining (16), (17) and (18) we obtain that

\[
\mathbb{E}|\hat{\Theta}_t - \Theta_t|^2 \\
\leq 2 \int_0^t \left( 2L^2 \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_{\lfloor \frac{s}{\eta} \rfloor}\right|^2 + 2L^2 \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_s \right|^2 \right) ds \\
+ 2\eta \int_0^t \left( 2L^2 \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_{\lfloor \frac{s}{\eta} \rfloor}\right|^2 + 2L^2 \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_s \right|^2 \right) ds. 
\]

\[
\mathbb{E}|\hat{\Theta}_t - \Theta_t|^2 \\
\leq 4(1 + \eta)L^2 \cdot \left( \int_0^t \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_{\lfloor \frac{s}{\eta} \rfloor}\right|^2 ds + \int_0^t \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_s \right|^2 ds \right). 
\]

Since we assumed that there is an \( M > 0 \) such that \( \max_{i=1,2,\ldots,N} (|\nabla_\theta \ell_i(\Theta)|) \leq M \) and \( |C(\Theta)| \leq \frac{1}{\sqrt{bN}} \sqrt{\sum_{i=1}^N |\nabla_\theta \ell_i(\Theta)|^2} \leq M \) since \( b \geq 1 \). By (10), the Itô’s isometry (Øksendal, 2003), the Cauchy-Schwarz inequality and \( 0 \leq s - \lfloor \frac{s}{\eta} \rfloor \eta \leq \eta \) we know that

\[
\mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_s \right|^2 \\
\leq \mathbb{E} \left| - \int_{\lfloor \frac{s}{\eta} \rfloor}^s \nabla_\theta L(\Theta_u) du + \sqrt{\eta} \int_{\lfloor \frac{s}{\eta} \rfloor}^s C(\Theta_u) dW_u \right|^2 \\
\leq 2\mathbb{E} \left( \int_{\lfloor \frac{s}{\eta} \rfloor}^s |\nabla_\theta L(\Theta_u)| du \right)^2 + 2\eta \mathbb{E} \left( \int_{\lfloor \frac{s}{\eta} \rfloor}^s C(\Theta_u) dW_u \right)^2 \\
\leq 2\eta \int_{\lfloor \frac{s}{\eta} \rfloor}^s \mathbb{E}|\nabla_\theta L(\Theta_u)|^2 du + 2\eta \int_{\lfloor \frac{s}{\eta} \rfloor}^s \mathbb{E}|C(\Theta_u)|^2 du \\
\leq 2\eta M^2 + 2\eta^2 M^2 = 4\eta^2 M^2. 
\]

Combining (20) and (19) we obtain

\[
\mathbb{E}|\hat{\Theta}_t - \Theta_t|^2 \leq 4(1 + \eta)L^2 \cdot \left( \int_0^t \mathbb{E}\left| \Theta_{\lfloor \frac{s}{\eta} \rfloor} - \Theta_{\lfloor \frac{s}{\eta} \rfloor}\right|^2 ds + 4\eta^2 M^2 t \right). 
\]

Set \( T > 0 \) and \( m(t) = \max_{0 \leq s \leq t} \mathbb{E}|\hat{\Theta}_s - \Theta_s|^2 \), noticing that \( m(\lfloor \frac{s}{\eta} \rfloor \eta) \leq m(s) \) (as \( \lfloor \frac{s}{\eta} \rfloor \eta \leq s \)), then the above gives for any \( 0 \leq t \leq T \),

\[
m(t) \leq 4(1 + \eta)L^2 \cdot \left( \int_0^t m(s) ds + 4\eta^2 M^2 t \right). 
\]
By Gronwall’s inequality we obtain that for $0 \leq t \leq T$,
\[
m(t) \leq 16(1 + \eta) L^2 \eta^2 M^2 T e^{4(1+\eta)L^2 t}.
\] (23)

Suppose $0 < \eta < 1$, then there is a constant $C$ which is independent on $\eta$ s.t.
\[
E|\hat{\Theta}_t - \Theta_t|^2 \leq m(t) \leq C \eta^2.
\] (24)

Set $t = k\eta$ in (24) and make use of (14), we finish the proof.

\[\square\]

**Remark.** As we have seen in the previous proof, the functions $\nabla_\theta L(\theta)$ and $C(\theta)$ are both $L$–Lipschitz continuous, and thus the SDE (10) admits a unique solution ((Øksendal, 2003), Section 5.2).

### C. Experiments setups and further results

The experiments are conducted using GeForce GTX 1080 Ti and PyTorch 1.0.0.

#### C.1. FashionMNIST

**Dataset**  [https://github.com/zalandoresearch/fashion-mnist](https://github.com/zalandoresearch/fashion-mnist)

We randomly choose 1,000 original test data as our training set, and use the 60,000 original training data as our test set. Thus we have 1,000 training data and 60,000 test data. We scale the image data to $[0, 1]$.

**Model**  We use a LeNet alike convolutional network:

\[
\text{input} \Rightarrow \text{conv1} \Rightarrow \text{max\_pool} \Rightarrow \text{ReLU} \Rightarrow \text{conv2} \Rightarrow \\
\text{max\_pool} \Rightarrow \text{ReLU} \Rightarrow \text{fc1} \Rightarrow \text{ReLU} \Rightarrow \text{fc2} \Rightarrow \text{output}.
\]

Both convolutional layers use $5 \times 5$ kernels with 10 channels and no padding. The number of hidden units between fully connected layers are 50. The total number of parameters of this network are 11,330.

**Optimization**  We use standard (stochastic) gradient descent optimizer. The learning rate is 0.01. If not stated otherwise, the batch size of SGD is 50.

#### C.2. SVHN

**Dataset**  [http://ufldl.stanford.edu/housenumbers/](http://ufldl.stanford.edu/housenumbers/)

We randomly choose 25,000 original test data as our training set, and 75,000 original training data as our test set. Thus we have 25,000 training data and 75,000 test data. We scale the image data to $[0, 1]$.

**Model**  We use standard VGG-11 without Batch Normalization.

**Optimization**  We use standard (stochastic) gradient descent optimizer. The learning rate is 0.05. If not stated otherwise, the batch size of SGD is 100.

#### C.3. CIFAR-10

**Dataset**  [https://www.cs.toronto.edu/~kriz/cifar.html](https://www.cs.toronto.edu/~kriz/cifar.html)

We use standard CIFAR-10 dataset. We scale the image into $[0, 1]$.

**Models**  We use two models: VGG-11 without Batch Normalization and standard ResNet-18.

**Optimization for VGG-11**  We use momentum (stochastic) gradient descent optimizer. The momentum is 0.9. The learning rate is 0.01 decayed by 0.1 at iteration 40,000 and 60,000. If not stated otherwise, the batch size of SGD is 100.
**Figure 3.** The generalization of MSGD. X-axis: number of iterations; y-axis: test accuracy. (a): We randomly draw 1,000 samples from FashionMNIST as the training set, then train a small convolutional network with them. (b): We use 25,000 samples from SVHN as the training set, then train a VGG-11 without Batch Normalization. **MSGD-Cov**: MSGD with Gaussian gradient noise whose covariance is the SGD covariance. **MSGD-[Cov-B]**: MSGD-Cov with the SGD covariance estimated using a mini-batch of samples in size $B$.

**Table 1.** Additional experiments for CIFAR-100 on ResNet-18

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD-500</td>
<td>76.38%</td>
</tr>
<tr>
<td>SGD-2k</td>
<td>72.78%</td>
</tr>
<tr>
<td>[MSGD-Fisher]-2k</td>
<td>76.83%</td>
</tr>
<tr>
<td>SGD-5k</td>
<td>59.16%</td>
</tr>
<tr>
<td>[MSGD-Fisher]-5k</td>
<td>76.46%</td>
</tr>
</tbody>
</table>

**Optimization for ResNet-18** We use momentum (stochastic) gradient descent optimizer. The momentum is 0.9. The learning rate is 0.1 decayed by 0.1 at iteration 40,000 and 60,000. If not stated otherwise, the batch size of SGD is 100. For large batch training, we use ghost batch normalization (Hoffer et al., 2017).

Specially, for the experiments to obtain state-of-the-art performance on ResNet-18, we also use standard data augmentation and weight decay $5 \times 10^{-4}$.

**C.4. Additional experiments**

**FashionMNIST and SVHN** Figure 3 shows additional experiments for MSGD-Cov. We see that indeed for MSGD-Cov, 1) the performance is similar to MSGD-Fisher, and 2) noises from different classes can generalize similarly.

**VGG-11** Figure 4 repeats our experiments in main text on VGG-11. The results are consistent with our main conclusions.

**CIFAR-100** Table 1 show additional result for CIFAR-100 on ResNet-18. The setups follow Figure 2 (c), except that the dataset is CIFAR-100 instead of CIFAR-10.
Figure 4. The generalization of MSGD and mini-batch MSGD. X-axis: number of iterations; y-axis: test accuracy. (a) (b): We train a VGG-11 on CIFAR-10 without using Batch Normalization, data augmentation and weight decay. **MSGD-Fisher**: MSGD with Gaussian gradient noise whose covariance is the scaled Fisher. **MSGD-Cov**: MSGD with Gaussian gradient noise whose covariance is the SGD covariance. **MSGD-Bernoulli**: MSGD with Bernoulli sampling noise. **SGD-B**: SGD with batch size $B$. **[MSGD-Fisher]-B**: mini-batch MSGD with batch size $B$, and an compensatory gradient noise whose covariance is the estimated Fisher. **[MSGD-Cov]-B**: mini-batch MSGD with batch size $B$, and an compensatory gradient noise whose covariance is the estimated SGD covariance.