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 $\mathcal{W}_{\rm sgd}$, then obtain that of the sampling noise $\mathcal{V}_{\rm sgd}$.

Sampling with replacement In the circumstance of sampling with replacement, the sampling random vector $\mathcal{W}_{\rm sgd}$ could be decompose as

$$\mathcal{W}_{\rm sgd} = \mathcal{W}^1 + \cdots + \mathcal{W}^b,$$

where $\mathcal{W}^1, \ldots, \mathcal{W}^b$ are i.i.d. and each of them represents once sampling procedure. Thus $\mathcal{W}^i = (w_1^i, \ldots, w_n^i)^T$ contains one multiple of $\frac{1}{b}$ and $n-1$ multiples of zero, with random index. Hence we have

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

Thus

$$\begin{aligned}
\mathbb{E}[\mathcal{W}^i] =& \frac{1}{bn}\mathbb{1}, \\
\mathrm{Var}[\mathcal{W}^i] =& \mathbb{E}[\mathcal{W}^i(\mathcal{W}^i)^T] - \mathbb{E}[\mathcal{W}^i]\mathbb{E}[\mathcal{W}^i]^T = \begin{pmatrix} \frac{1}{b^2 n} & & \\ & \ddots & \\ & & \frac{1}{b^2 n} \end{pmatrix} - \frac{1}{b^2 n^2}\mathbb{1}\mathbb{1}^T = \frac{1}{b^2 n}\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right).
\end{aligned}$$

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

$$\mathbb{E}[\mathcal{W}_{\rm sgd}] = b\mathbb{E}[\mathcal{W}^i] = \frac{1}{n}\mathbb{1}, \quad \mathrm{Var}[\mathcal{W}_{\rm sgd}] = b\,\mathrm{Var}[\mathcal{W}^i] = \frac{1}{bn}\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right).$$

Therefore for the sampling noise $\mathcal{V}_{\rm sgd} = \mathcal{W}_{\rm sgd} - \frac{1}{n}\mathbb{1}$ we have

$$\mathbb{E}[\mathcal{V}_{\rm sgd}] = 0, \quad \mathrm{Var}[\mathcal{V}_{\rm sgd}] = \frac{1}{bn}\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right).$$

Sampling without replacement Let $\mathcal{W}'_{\rm sgd} = (w'_1, \ldots, w'_n)^T$. In the case of sampling without replacement, we know the sampling random vector $\mathcal{W}'_{\rm sgd}$ contains exactly b multiples of $\frac{1}{b}$s and $n-b$ multiples of zero, with random index. Hence we have

$$\mathbb{E}[w'_j] = \frac{\binom{n-1}{b-1}\frac{1}{b}}{\binom{n}{b}} = \frac{1}{n}, \quad \mathbb{E}[(w'_j)^2] = \frac{\binom{n-1}{b-1}\frac{1}{b^2}}{\binom{n}{b}} = \frac{1}{bn}, \quad \mathbb{E}[w'_j w'_k] = \frac{\binom{n-2}{b-2}\frac{1}{b^2}}{\binom{n}{b}} = \frac{b-1}{bn(n-1)},\ \forall j \neq k.$$

Thus

$$\begin{aligned}
\mathbb{E}[\mathcal{W}'_{\rm sgd}] =& \frac{1}{N}\mathbb{1}, \\
\mathrm{Var}[\mathcal{W}'_{\rm sgd}] =& \mathbb{E}[\mathcal{W}'_{\rm sgd}(\mathcal{W}'_{\rm sgd})^T] - \mathbb{E}[\mathcal{W}^i]\mathbb{E}[\mathcal{W}^i]^T \\
=& \begin{pmatrix} \frac{1}{bn} & \frac{b-1}{bn(n-1)} & \cdots & \frac{b-1}{bn(n-1)} \\ \frac{b-1}{bn(n-1)} & \frac{1}{bn} & \cdots & \frac{b-1}{bn(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b-1}{bn(n-1)} & \frac{b-1}{bn(n-1)} & \cdots & \frac{1}{b^2 n} \end{pmatrix} - \frac{1}{n^2}\mathbb{1}\mathbb{1}^T = \frac{n-b}{bn(n-1)}\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right).
\end{aligned}$$

Therefore for the sampling noise $\mathcal{V}'_{\rm sgd} = \mathcal{W}'_{\rm sgd} - \frac{1}{n}\mathbb{1}$ we have

$$\mathbb{E}[\mathcal{V}'_{\rm sgd}] = 0, \quad \mathrm{Var}[\mathcal{V}'_{\rm sgd}] = \frac{n-b}{bn(n-1)}\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right).$$

□

A.2. Proof of Theorem 1

Proof. Let

$$\epsilon_n = y_n - x_n^T \theta_*,$$

by assumption we have

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

Recall the MSGD updates

$$\theta_{n+1} = \theta_n - \eta \sum_{r \in B_n} w_r \left(x_r x_r^T \theta_n - y_r x_r \right),$$

hence we have

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

Define

$$L(k) = \sum_{r \in B_k} w_r x_r x_r^T, \tag{5}$$

$$M(i,k) = \begin{cases} (I - \eta L(i)) \cdots (I - \eta L(k)), & i \geq k \\ I, & i < k. \end{cases} \tag{6}$$

$$N(k) = \sum_{r \in B_k} w_r \epsilon_r x_r, \tag{7}$$

Then recursively we obtain

$$\theta_i - \theta_* = M(i,1)(\theta_0 - \theta_*) + \eta \sum_{k=1}^{i} M(i,k+1) N(k). \tag{8}$$

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

$$\begin{aligned}
\mathbb{E}\left[L(k)\right] =& \sum_{r=1}^{B} \mathbb{E}[w_r] \cdot \mathbb{E}[x_r x_r^T] = B \cdot \frac{1}{B} \cdot \Sigma = \Sigma. \\
\mathbb{E}\left[L(k)^2\right] =& \mathbb{E} \sum_{r=1}^{B} w_r^2 x_r x_r^T x_r x_r^T + 2\mathbb{E} \sum_{r=1}^{B-1} \sum_{s=2}^{B} w_r w_s x_r x_r^T x_s x_s^T \\
=& \sum_{r=1}^{B} \mathbb{E}[w_r^2] \cdot \mathbb{E}\left[\|x_r\|_2^2 x_r x_r^T \right] + 2 \sum_{r=1}^{B-1} \sum_{s=2}^{B} \mathbb{E}[w_r w_s] \cdot \mathbb{E}[x_r x_r^T] \cdot \mathbb{E}[x_s x_s^T] \\
=& B \cdot \frac{1}{bB} \cdot \mathbb{E}\left[\|x\|_2^2 x x^T \right] + 2 \frac{B(B-1)}{2} \cdot \frac{b-1}{bB(B-1)} \cdot \Sigma^2 \\
\preceq& \frac{1}{b} R^2 \Sigma + \frac{b-1}{b} \lambda \Sigma = \frac{R^2 + (b-1)\lambda}{b} \Sigma.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[M(i,k)] =& (I - \eta \mathbb{E}[L(i)]) \cdots (I - \eta \mathbb{E}[L(k)]) = (I - \eta \Sigma)^{i-k+1}. \\
\mathbb{E} M(i,k) M(i,k)^T =& \mathbb{E} M(i,k+1) \left(I - \eta L(k) \right)^2 M(i,k+1)^T \\
=& \mathbb{E} M(i,k+1) \left(I - 2\eta L(k) + \eta^2 L(k)^2 \right) M(i,k+1)^T \\
\leq& \mathbb{E} M(i,k+1) \left(I - 2\eta \Sigma + \eta^2 \frac{R^2 + (b-1)\lambda}{b} \Sigma \right) M(i,k+1)^T \\
=& \mathbb{E} M(i,k+1) M(i,k+1)^T - \eta \left(2 - \eta \frac{R^2 + (b-1)\lambda}{b} \right) \mathbb{E} M(i,k+1) \Sigma M(i,k+1)^T.
\end{aligned}$$

Hence

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

Moments of $N(k)$

$$\begin{aligned}
\mathbb{E}[N(k)] =& \sum_{r\in B_k}\mathbb{E}[w_r]\cdot\mathbb{E}[\epsilon_r x_r] = B\cdot\frac{1}{B}\cdot 0 = 0.\\
\mathbb{E}\left[N(k)N(k)^T\right] =& \mathbb{E}\sum_{r=1}^{B} w_r^2\epsilon_r^2 x_r x_r^T + 2\mathbb{E}\sum_{r=1}^{B-1}\sum_{s=2}^{B} w_r w_s \epsilon_r\epsilon_s x_r x_s^T\\
=& \sum_{r=1}^{B}\mathbb{E}[w_r^2]\cdot\mathbb{E}[\epsilon_r^2 x_r x_r^T] + 2\sum_{r=1}^{B-1}\sum_{s=2}^{B}\mathbb{E}[w_r]\cdot\mathbb{E}[w_s]\cdot\mathbb{E}[\epsilon_r x_r]\cdot\mathbb{E}[\epsilon_s x_s]^T\\
=& B\cdot\frac{1}{bB}\cdot\mathbb{E}[\epsilon^2 x x^T] + 2\frac{B(B-1)}{2}\cdot\frac{1}{B^2}\cdot 0\\
\leq& \frac{1}{b}\sigma^2\Sigma.
\end{aligned}$$

Calculate averaging Takeing expectation to w_k and B_k, we have

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle\\
=&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\left\langle\theta_i-\theta_*,\Sigma\left(M(j,i+1)(\theta_i-\theta_*)+\eta\sum_{k=i+1}^{j}M(j,k+1)N(k)\right)\right\rangle\\
=&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma M(j,i+1)(\theta_i-\theta_*)\rangle\\
=&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\left\langle\theta_i-\theta_*,\Sigma\left(I-\eta\Sigma\right)^{j-i}(\theta_i-\theta_*)\right\rangle\\
=&\mathbb{E}\sum_{i=0}^{n-1}\left\langle\theta_i-\theta_*,\eta^{-1}\left(I-\eta\Sigma-(I-\eta\Sigma)^{n-i+1}\right)(\theta_i-\theta_*)\right\rangle\\
\leq&\mathbb{E}\sum_{i=0}^{n}\left\langle\theta_i-\theta_*,\eta^{-1}\left(I-\eta\Sigma\right)(\theta_i-\theta_*)\right\rangle\\
=&\eta^{-1}\mathbb{E}\sum_{i=0}^{n}\|\theta_i-\theta_*\|_2^2-\mathbb{E}\sum_{i=0}^{n}\left\|\Sigma^{\frac{1}{2}}(\theta_i-\theta_*)\right\|_2^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
&(n+1)^2\mathbb{E}\left\|\Sigma^{\frac{1}{2}}(\bar{\theta}_n-\theta_*)\right\|_2^2 = \mathbb{E}\sum_{i,j=0}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle\\
=&\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_i-\theta_*)\rangle + 2\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle\\
\leq&\mathbb{E}\sum_{i=0}^{n}\left\|\Sigma^{\frac{1}{2}}(\theta_i-\theta_*)\right\|_2^2 + 2\eta^{-1}\mathbb{E}\sum_{i=0}^{n}\|\theta_i-\theta_*\|_2^2 - 2\mathbb{E}\sum_{i=0}^{n}\left\|\Sigma^{\frac{1}{2}}(\theta_i-\theta_*)\right\|_2^2\\
\leq&2\eta^{-1}\mathbb{E}\sum_{i=0}^{n}\|\theta_i-\theta_*\|_2^2,
\end{aligned}$$

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

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

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

In conclusion we have

$$\begin{aligned}&\frac{1}{2}\eta(n+1)^2\mathbb{E}\left\|\Sigma^{\frac{1}{2}}(\bar{\theta}_n-\theta_*)\right\|_2^2\le\mathbb{E}\sum_{i=0}^{n}\|\theta_i-\theta_*\|_2^2\\ =&\mathbb{E}\sum_{i=0}^{n}\|M(i,1)(\theta_0-\theta_*)\|_2^2+\eta^2\mathbb{E}\sum_{i=0}^{n}\sum_{k=1}^{i}\langle M(i,k+1)N(k),\ M(i,k+1)N(k)\rangle\,.\end{aligned}$$

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

Noise term We bound the noise term by observing that

$$\begin{aligned}&\mathbb{E}\langle M(i,k+1)N(k),\ M(i,k+1)N(k)\rangle\\ =&\mathbb{E}M(i,k+1)N(k)N(k)^TM(i,k+1)^T\\ =&\mathbb{E}\operatorname{Tr}\left[N(k)N(k)^TM(i,k+1)^TM(i,k+1)\right]\\ =&\operatorname{Tr}\left[\mathbb{E}\left[N(k)N(k)^T\right]\cdot\mathbb{E}\left[M(i,k+1)^TM(i,k+1)\right]\right]\\ \preceq&\sigma^2\operatorname{Tr}\left[\Sigma\cdot\mathbb{E}\left[M(i,k+1)^TM(i,k+1)\right]\right]\\ =&\sigma^2\operatorname{Tr}\mathbb{E}\left[M(i,k+1)\Sigma M(i,k+1)^T\right]\\ \le&\frac{\sigma^2}{\eta\left(2-\eta\frac{R^2+(b-1)\lambda}{b}\right)}\operatorname{Tr}\left(\mathbb{E}M(i,k+1)M(i,k+1)^T-\mathbb{E}M(i,k)M(i,k)^T\right).\end{aligned}$$

Hence

$$\begin{aligned}&\eta^2\mathbb{E}\sum_{i=0}^{n}\sum_{k=1}^{i}\langle M(i,k+1)N(k),\ M(i,k+1)N(k)\rangle\\ \le&\frac{\eta\sigma^2}{\left(2-\eta\frac{R^2+(b-1)\lambda}{b}\right)}\sum_{i=0}^{n}\sum_{k=1}^{i}\operatorname{Tr}\left(\mathbb{E}M(i,k+1)M(i,k+1)^T-\mathbb{E}M(i,k)M(i,k)^T\right)\\ =&\frac{\eta\sigma^2}{\left(2-\eta\frac{R^2+(b-1)\lambda}{b}\right)}\sum_{i=0}^{n}\operatorname{Tr}\left(\mathbb{E}M(i,i+1)M(i,i+1)^T-\mathbb{E}M(i,1)M(i,1)^T\right)\\ \le&\frac{\eta\sigma^2}{\left(2-\eta\frac{R^2+(b-1)\lambda}{b}\right)}\sum_{i=0}^{n}\operatorname{Tr}\left(\mathbb{E}M(i,i+1)M(i,i+1)^T\right)\\ =&\frac{\eta\sigma^2}{\left(2-\eta\frac{R^2+(b-1)\lambda}{b}\right)}(n+1)\operatorname{Tr}[I]=\frac{\eta\sigma^2}{\left(2-\eta\frac{R^2+(b-1)\lambda}{b}\right)}(n+1)d.\end{aligned}$$

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

$$\begin{aligned}SA=&\mathbb{E}\left[L(k)AL(k)\right]\\ TA=&\Sigma A+A\Sigma-\eta\mathbb{E}\left[L(k)AL(k)\right]=\Sigma A+A\Sigma-\eta SA.\end{aligned}$$

With these notations and $M(i,1) = (I - \eta L(i)) \cdots (I - \eta L(1))$, we recursively have

$$\mathbb{E}\left[M(i,1)^T M(i,1)\right] = (I - \eta T)^i I.$$

Next we bound the noiseless term

$$\begin{aligned}
\mathbb{E}\sum_{i=0}^{n} \|M(i,1)(\theta_0 - \theta_*)\|_2^2 =& \mathbb{E}\sum_{i=0}^{n} \operatorname{Tr}\left[M(i,1)^T M(i,1)(\theta_0 - \theta_*)(\theta_0 - \theta_*)^T\right] \\
=& \sum_{i=0}^{n} \left\langle \mathbb{E} M(i,1)^T M(i,1), E_0 \right\rangle \\
=& \sum_{i=0}^{n} \left\langle (I - \eta T)^i I, E_0 \right\rangle \\
=& \left\langle \eta^{-1} T^{-1} \left(I - (I - \eta T)^{n+1}\right) I, E_0 \right\rangle \\
\leq& \left\langle \eta^{-1} T^{-1} I, E_0 \right\rangle.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{n} \|M(i,1)(\theta_0 - \theta_*)\|_2^2 = \left\langle \eta^{-1} M, E_0 \right\rangle = \frac{1}{2\eta} \left\langle \Sigma^{-1}, E_0 \right\rangle + \left\langle (\Sigma \otimes I + I \otimes \Sigma)^{-1} SM, E_0 \right\rangle \\
=& \frac{1}{2\eta} (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) + \left\langle SM, (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \right\rangle.
\end{aligned}$$

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

$$E_0 = \Sigma^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\theta_0 - \theta_*)(\theta_0 - \theta_*)^T \Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}} \preceq (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \Sigma.$$

Thus

$$(\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \preceq (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot (\Sigma \otimes I + I \otimes \Sigma)^{-1} \Sigma = (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \frac{1}{2} I.$$

Bound SM Firstly by definition,

$$\begin{aligned}
\operatorname{Tr}[SM] =& \mathbb{E} \operatorname{Tr}\left[L(k) M L(k)\right] = \mathbb{E}\sum_{r=1}^{B} \operatorname{Tr}[w_r^2 x_r x_r^T M x_r x_r^T] + 2\mathbb{E}\sum_{r=1}^{B-1}\sum_{s=2}^{B} \operatorname{Tr}[w_r w_s x_r x_r^T M x_s x_s^T] \\
=& \sum_{r=1}^{B} \mathbb{E}[w_r^2] \cdot \operatorname{Tr}\left[\mathbb{E}\left[\|x_r\|_2^2 x_r x_r^T\right] M\right] + 2\sum_{r=1}^{B-1}\sum_{s=2}^{B} \mathbb{E}[w_r w_s] \cdot \operatorname{Tr}\left[\mathbb{E}[x_r x_r^T] \cdot M \cdot \mathbb{E}[x_s x_s^T]\right] \\
\leq& B \cdot \frac{1}{bB} \cdot \operatorname{Tr}\left[R^2 \Sigma M\right] + 2\frac{B(B-1)}{2} \cdot \frac{b-1}{bB(B-1)} \cdot \operatorname{Tr}\left[\Sigma M \Sigma\right] \\
\leq& \frac{R^2}{b} \cdot \operatorname{Tr}\left[\Sigma M\right] + \frac{b-1}{b} \cdot \lambda \cdot \operatorname{Tr}\left[M\Sigma\right] = \frac{R^2 + (b-1)\lambda}{b} \operatorname{Tr}\left[\Sigma M\right].
\end{aligned}$$

Secondly taking trace we have

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

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

To sum up we have

$$\begin{aligned}
&\left\langle SM,\ (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \right\rangle \\
\leq &\frac{1}{2}(\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \langle SM, I \rangle \\
= &\frac{1}{2}(\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) \cdot \operatorname{Tr}[SM] \\
\leq &\frac{(R^2 + (b-1)\lambda)d}{4b} (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*).
\end{aligned}$$

Therefore for the noiseless term we have

$$\begin{aligned}
&\mathbb{E} \sum_{i=0}^{n} \|M(i,1)(\theta_0 - \theta_*)\|_2^2 = \frac{1}{2\eta}(\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*) + \left\langle SM, (\Sigma \otimes I + I \otimes \Sigma)^{-1} E_0 \right\rangle \\
\leq &\left(\frac{1}{2\eta} + \frac{(R^2 + (b-1)\lambda)d}{4b} \right) (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*).
\end{aligned}$$

In conclusion we have

$$\begin{aligned}
&\frac{1}{2}\eta(n+1)^2 \mathbb{E} \left\| \Sigma^{\frac{1}{2}} (\bar{\theta}_n - \theta_*) \right\|_2^2 \leq \text{noiseless term} + \text{noise term} \\
\leq &\frac{\eta \sigma^2}{\left(2 - \eta \frac{R^2+(b-1)\lambda}{b}\right)} (n+1)d + \left(\frac{1}{2\eta} + \frac{(R^2 + (b-1)\lambda)d}{4b} \right) (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*).
\end{aligned}$$

Hence

$$\mathbb{E} \left\| \Sigma^{\frac{1}{2}} (\bar{\theta}_n - \theta_*) \right\|_2^2 \leq \frac{1}{n+1} \cdot \frac{2\sigma^2 d}{\left(2 - \eta \frac{R^2+(b-1)\lambda}{b}\right)} + \frac{1}{(n+1)^2} \cdot \left(1 + \frac{(R^2 + (b-1)\lambda)\eta d}{2b} \right) (\theta_0 - \theta_*)^T \Sigma^{-1} (\theta_0 - \theta_*),$$

which complete our proof.

□

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 $\max_{i=1,2,\ldots,N} (|\nabla_\theta \ell_i(\theta)|) \leq M$ and that $\nabla \ell_i(\theta)$ are Lipschitz continuous with bounded Lipschitz constant $L > 0$ uniformly for all $i = 1, 2, ..., N$.*

Then the Gaussian MSGD iteration (9)

$$\theta_{k+1} - \theta_k = -\eta \nabla_\theta L(\theta_k) + \eta C(\theta_k) \mathcal{W}_{k+1},\ \mathcal{W}_k \sim \mathcal{N}(0, I),\ i.i.d. \tag{9}$$

is a order 1 *strong approximation to SDE* (10)

$$\mathrm{d}\Theta_t = -\nabla_\theta L(\Theta_t)\mathrm{d}t + \sqrt{\eta} C(\Theta_t)\mathrm{d}W_t,\ \Theta_0 = \theta_0,\ W_t \in \mathbb{R}^N \text{ is a standard Brownian motion} \tag{10}$$

i.e., there exist a constant C independent on η but depending on L and M such that

$$\mathbb{E}\|\Theta_{k\eta} - \theta_k\|^2 \leq C\eta^2, \quad \text{for all } 0 \leq k \leq \lfloor T/\eta \rfloor. \tag{11}$$

Proof. We show that, as $\eta \to 0$, the discrete iteration θ_k 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 $|X| := \sqrt{\mathrm{Tr}(X^T X)} = \sqrt{\mathrm{Tr}(X X^T)}$.

Let $\widehat{\Theta}_t$ be the process defined by the integral form of the stochastic differential equation

$$\widehat{\Theta}_t - \widehat{\Theta}_0 = -\int_0^t \nabla_\theta L(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta}) \mathrm{d}s + \sqrt{\eta} \int_0^t C(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta}) \mathrm{d}W_s \ , \ \widehat{\Theta}_0 = \theta_0 \ . \tag{12}$$

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

$$\widehat{\Theta}_{(k+1)\eta} - \widehat{\Theta}_{k\eta} = -\eta \nabla_\theta L(\widehat{\Theta}_{k\eta}) - \sqrt{\eta} C(\widehat{\Theta}_{k\eta})(W_{(k+1)\eta} - W_{k\eta}) \ . \tag{13}$$

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

$$\widehat{\Theta}_{k\eta} = \theta_k \ , \tag{14}$$

where θ_k is the solution to (9).

We first bound $\widehat{\Theta}_t$ in Eq. (12) and Θ_t in Eq. (10). Then we could obtain the error estimation of $\theta_k = \widehat{\Theta}_{k\eta}$ 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 $|C(\theta_1) - C(\theta_2)| = \frac{1}{\sqrt{bN}} \sqrt{\sum_{i=1}^N |\nabla_\theta \ell_i(\theta_1) - \nabla_\theta \ell_i(\theta_2)|^2} \leq \frac{1}{\sqrt{bN}} \sqrt{NL^2 |\theta_1 - \theta_2|^2} \leq L|\theta_1 - \theta_2|$ since $b \geq 1$. Thus $C(\theta)$ is also L–Lipschitz continuous. Take a difference between (12) and (10) we get

$$\widehat{\Theta}_t - \Theta_t = -\int_0^t [\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta}) - \nabla_\theta L(\Theta_s)] \mathrm{d}s + \sqrt{\eta} \int_0^t [C(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor}) - C(\Theta_s)] \mathrm{d}W_s \ . \tag{15}$$

We can estimate

$$\begin{aligned} &|\nabla_\theta L(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta}) - \nabla_\theta L(\Theta_s)|^2 \\ \leq & 2|\nabla_\theta L(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta}) - \nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta})|^2 + 2|\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta}) - \nabla_\theta L(\Theta_s)|^2 \\ \leq & 2L^2 |\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta} - \Theta_{\lfloor \frac{s}{\eta} \rfloor \eta}|^2 + 2L^2 |\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta} - \Theta_s|^2 \ , \end{aligned} \tag{16}$$

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

$$\begin{aligned} &|C(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta}) - C(\Theta_s)|^2 \\ \leq & 2|C(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta}) - C(\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta})|^2 + 2|C(\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta}) - C(\Theta_s)|^2 \\ \leq & 2L^2 |\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta} - \Theta_{\lfloor \frac{s}{\eta} \rfloor \eta}|^2 + 2L^2 |\Theta_{\lfloor \frac{s}{\eta} \rfloor \eta} - \Theta_s|^2 \ . \end{aligned} \tag{17}$$

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

$$\begin{aligned}
&\mathbb{E}|\widehat{\Theta}_t - \Theta_t|^2 \\
\leq& 2\mathbb{E}\left|\int_0^t [\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}) - \nabla_\theta L(\Theta_s)]\mathrm{d}s\right|^2 + 2\eta\mathbb{E}\left|\int_0^t [C(\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor}) - C(\Theta_s)]\mathrm{d}W_s\right|^2 \\
\leq& 2\mathbb{E}\left|\int_0^t [\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}) - \nabla_\theta L(\Theta_s)]\mathrm{d}s\right|^2 + 2\eta\int_0^t \mathbb{E}\left|C(\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor}) - C(\Theta_s)\right|^2 \mathrm{d}s \\
\leq& 2\int_0^t \mathbb{E}\left|\nabla_\theta L(\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}) - \nabla_\theta L(\Theta_s)\right|^2 \mathrm{d}s + 2\eta\int_0^t \mathbb{E}\left|C(\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor}) - C(\Theta_s)\right|^2 \mathrm{d}s\,.
\end{aligned} \tag{18}$$

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

$$\begin{aligned}
&\mathbb{E}|\widehat{\Theta}_t - \Theta_t|^2 \\
\leq& 2\int_0^t \left(2L^2\mathbb{E}|\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}|^2 + 2L^2\mathbb{E}|\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_s|^2\right)\mathrm{d}s \\
&+2\eta\int_0^t \left(2L^2\mathbb{E}|\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}|^2 + 2L^2\mathbb{E}|\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_s|^2\right)\mathrm{d}s\,. \\
=& 4(1+\eta)L^2\cdot\left(\int_0^t \mathbb{E}|\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}|^2\mathrm{d}s + \int_0^t \mathbb{E}|\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_s|^2\mathrm{d}s\right)\,.
\end{aligned} \tag{19}$$

Since we assumed that there is an $M > 0$ such that $\max_{i=1,2\ldots,N}(|\nabla_\theta \ell_i(\theta)|) \leq M$, we conclude that $|\nabla_\theta L(\theta)| \leq \frac{1}{N}\sum_{i=1}^N |\nabla_\theta \ell_i(\theta)| \leq M$ 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

$$\begin{aligned}
&\mathbb{E}|\Theta_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_s|^2 \\
=& \mathbb{E}\left|-\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s \nabla_\theta L(\Theta_u)\mathrm{d}u + \sqrt{\eta}\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s C(\Theta_u)\mathrm{d}W_u\right|^2 \\
\leq& 2\mathbb{E}\left|\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s \nabla_\theta L(\Theta_u)\mathrm{d}u\right|^2 + 2\eta\mathbb{E}\left|\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s C(\Theta_u)\mathrm{d}W_u\right|^2 \\
\leq& 2\mathbb{E}\left(\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s |\nabla_\theta L(\Theta_u)|\,\mathrm{d}u\right)^2 + 2\eta\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s \mathbb{E}|C(\Theta_u)|^2\mathrm{d}u \\
\leq& 2\eta\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s \mathbb{E}|\nabla_\theta L(\Theta_u)|^2\mathrm{d}u + 2\eta\int_{\lfloor \frac{s}{\eta}\rfloor \eta}^s \mathbb{E}|C(\Theta_u)|^2\mathrm{d}u \\
\leq& 2\eta^2M^2 + 2\eta^2M^2 = 4\eta^2M^2\,.
\end{aligned} \tag{20}$$

Combining (20) and (19) we obtain

$$\mathbb{E}|\widehat{\Theta}_t - \Theta_t|^2 \leq 4(1+\eta)L^2\cdot\left(\int_0^t \mathbb{E}|\widehat{\Theta}_{\lfloor \frac{s}{\eta}\rfloor \eta} - \Theta_{\lfloor \frac{s}{\eta}\rfloor \eta}|^2\mathrm{d}s + 4\eta^2M^2t\right)\,. \tag{21}$$

Set $T > 0$ and $m(t) = \max_{0\leq s\leq t} \mathbb{E}|\widehat{\Theta}_s - \Theta_s|^2$, 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 $0 \leq t \leq T$,

$$m(t) \leq 4(1+\eta)L^2\cdot\left(\int_0^t m(s)\mathrm{d}s + 4\eta^2M^2T\right)\,. \tag{22}$$

By Gronwall's inequality we obtain that for $0 \leq t \leq T$,

$$m(t) \leq 16(1+\eta)L^2\eta^2M^2Te^{4(1+\eta)L^2t}. \tag{23}$$

Suppose $0 < \eta < 1$, then there is a constant C which is independent on η s.t.

$$\mathbb{E}|\widehat{\Theta}_t - \Theta_t|^2 \leq m(t) \leq C\eta^2. \tag{24}$$

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

□

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

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

C.1. FashionMNIST

Dataset `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:

$$\begin{aligned}\text{input} \Rightarrow \text{conv1} \Rightarrow \text{max\_pool} \Rightarrow \text{ReLU} \Rightarrow \text{conv2} \Rightarrow \\ \text{max\_pool} \Rightarrow \text{ReLU} \Rightarrow \text{fc1} \Rightarrow \text{ReLU} \Rightarrow \text{fc2} \Rightarrow \text{output}.\end{aligned}$$

Both convolutional layers use 5×5 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/`

We randomly choose $25,000$ original test data as our training set, and $70,000$ original training data as our test set. Thus we have $25,000$ training data and $70,000$ test data. We scale the image data to $[0, 1]$.

Model We use standard VGG-11 without Batch Normalization.

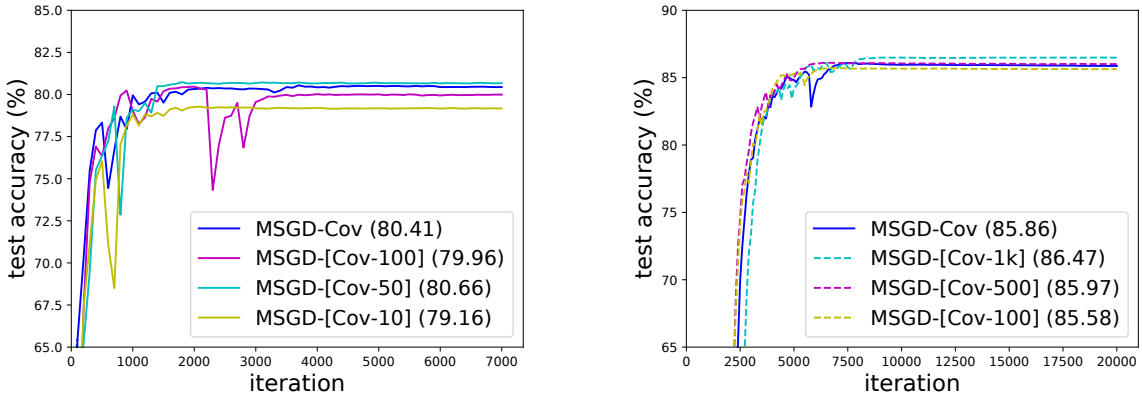
Optimization We use standard (stochastic) gradient descent optimizer. The learning rate is 0.05. If not stated otherwise, the batch size of SGD is 100.

C.3. CIFAR-10

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

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.



(a) Small FashionMNIST (b) SVHN

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. Additioal experimetns for CIFAR-100 on ResNet-18

Algorithm	Test Accuracy
SGD-500	76.38%
SGD-2k	72.78%
[MSGD-Fisher]-2k	76.83%
SGD-5k	59.16%
[MSGD-Fisher]-5k	76.46%

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.

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).

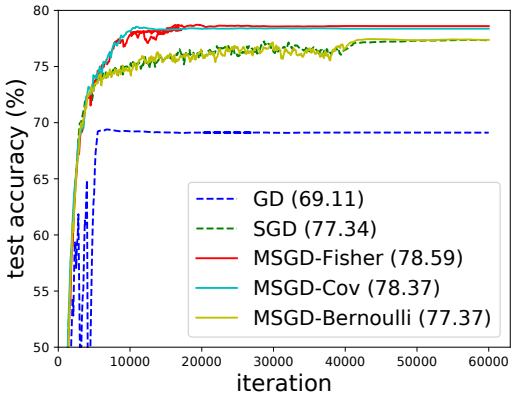
Specially, for the experiments to obtain state-of-the-art performance on ResNet-18, we also use standard data augmentation and weight decay 5×10^{-4}.

C.4. Additional experiments

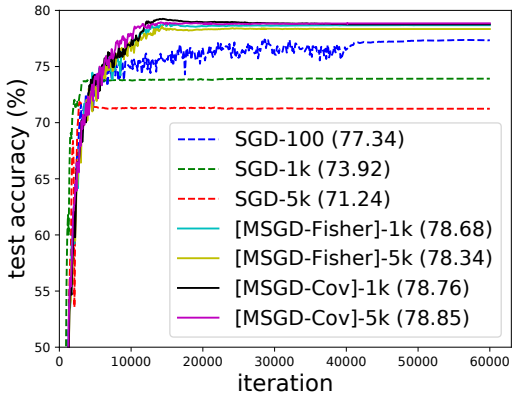
FashionMNIST and SVHN Figure 3 shows additional experiments for MSGD-Cov. We see that indeed for MSGD-Cov, 1) the performance is similar to MSGD-Fisher, and 2) noises from different classes can generalize similarly.

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 addtional result for CIFAR-100 on ResNet-18. The setups follow Figure 2 (c), except that the dataset is CIFAR-100 instead of CIFAR-10.



(a) CIFAR-10, VGG-11



(b) CIFAR-10, VGG-11

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. textbf[MSGD-Cov]-**B**: mini-batch MSGD with batch size B, and an compensatory gradient noise whose covariance is the estimated SGD covariance.