Communication trade-offs for synchronized distributed SGD with large step size

Aymeric DIEULEVEUT

EPLF, MLO

17 November 2017

Joint work with Kumar Kshitij Patel.
Outline

1. Stochastic gradient descent - supervised machine learning - setting, assumptions and proof techniques
2. Synchronized distributed SGD - from mini-batch averaging to model averaging
3. Optimality of Local-SGD.
Stochastic Gradient Descent

- **Goal:**
  \[
  \min_{\theta \in \mathbb{R}^d} F(\theta)
  \]
  given unbiased gradient estimates \( g_n \)

- \( \theta^* := \arg\min_{\mathbb{R}^d} F(\theta) \).
Stochastic Gradient Descent

Goal:

\[ \min_{\theta \in \mathbb{R}^d} F(\theta) \]

given unbiased gradient estimates \( g_n \)

\[ \theta^* := \text{argmin}_{\mathbb{R}^d} F(\theta). \]

Key algorithm: Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951):

\[ \theta_k = \theta_{k-1} - \eta_k \ g_k(\theta_{k-1}) \]

\[ \mathbb{E}[g_k(\theta_{k-1})|\mathcal{F}_{k-1}] = F'(\theta_{k-1}) \] for a filtration \( (\mathcal{F}_k)_{k \geq 0} \), \( \theta_k \) is \( \mathcal{F}_k \) measurable.
Stochastic Gradient Descent

Goal:
\[
\min_{\theta \in \mathbb{R}^d} F(\theta)
\]
given unbiased gradient estimates \(g_n\)

\(\theta^* := \arg\min_{\mathbb{R}^d} F(\theta)\).

Key algorithm: Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951):

\[
\theta_k = \theta_{k-1} - \eta_k g_k(\theta_{k-1})
\]

\[\mathbb{E}[g_k(\theta_{k-1})|\mathcal{F}_{k-1}] = F'(\theta_{k-1})\text{ for a filtration } (\mathcal{F}_k)_{k \geq 0}, \text{ } \theta_k \text{ is } \mathcal{F}_k \text{ measurable.}\]
We define the risk (generalization error) as

$$\mathcal{R}(\theta) := \mathbb{E}_\rho [\ell(Y, \langle \theta, \Phi(X) \rangle)].$$

Empirical risk (or training error):

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \Phi(x_i) \rangle).$$
Supervised Machine Learning

- We define the risk (generalization error) as
  \[ \mathcal{R}(\theta) := \mathbb{E}_\rho [\ell(Y, \langle \theta, \Phi(X) \rangle)]. \]

- Empirical risk (or training error):
  \[ \hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \Phi(x_i) \rangle). \]

- For example, least-squares regression:
  \[ \min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \theta, \Phi(x_i) \rangle)^2 + \mu \Omega(\theta), \]

- and logistic regression:
  \[ \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log (1 + \exp(-y_i \langle \theta, \Phi(x_i) \rangle)) + \mu \Omega(\theta). \]
Polyak Ruppert averaging
Polyak Ruppert averaging

Introduced by Polyak and Juditsky (1992) and Ruppert (1988):

$$\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^{n} \theta_k.$$ 

- off line averaging reduces the noise effect.
Polyak Ruppert averaging

Introduced by Polyak and Juditsky (1992) and Ruppert (1988):

\[ \bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^{n} \theta_k. \]

- off line averaging reduces the noise effect.
- on line computing: \( \bar{\theta}_{n+1} = \frac{1}{n+1} \theta_{n+1} + \frac{n}{n+1} \bar{\theta}_n. \)
Assumptions

Goal: \[ \min_{\theta} F(\theta) \]

Recursion: \[ \theta_k = \theta_{k-1} - \eta_k g_k(\theta_{k-1}) \]

A1 [Strong convexity] The function \( F \) is strongly-convex with convexity constant \( \mu > 0 \).
Assumptions

Goal: \( \min_{\theta} F(\theta) \). Recursion: \( \theta_k = \theta_{k-1} - \eta_k g_k(\theta_{k-1}) \)

A1 [Strong convexity] The function \( F \) is strongly-convex with convexity constant \( \mu > 0 \).

A2 [Smoothness and regularity] The function \( F \) is three times continuously differentiable with second and third uniformly bounded derivatives: \( \sup_{\theta \in \mathbb{R}^d} \| F^{(2)}(\theta) \| < L \), and \( \sup_{\theta \in \mathbb{R}^d} \| F^{(3)}(\theta) \| < M \). Especially \( F \) is \( L \)-smooth.
Assumptions

Goal: $\min_{\theta} F(\theta)$.  
Recursion: $\theta_k = \theta_{k-1} - \eta_k \, g_k(\theta_{k-1})$

A1 [Strong convexity] The function $F$ is strongly-convex with convexity constant $\mu > 0$.

A2 [Smoothness and regularity] The function $F$ is three times continuously differentiable with second and third uniformly bounded derivatives: $\sup_{\theta \in \mathbb{R}^d} \left\| F^{(2)}(\theta) \right\| < L$, and $\sup_{\theta \in \mathbb{R}^d} \left\| F^{(3)}(\theta) \right\| < M$. Especially $F$ is $L$-smooth. Or:

Q1 [Quadratic function] There exists a positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, such that the function $F$ is the quadratic function $\theta \mapsto \| \Sigma^{1/2}(\theta - \theta^*) \|^2 / 2$,.
Which step size would you use?

**Smooth functions.**

\[ \eta_k \equiv \eta_0 \quad \eta_k = \frac{1}{\sqrt{k}} \quad \eta_k = \frac{1}{(\mu k)} \]

- Convex
- Strongly Convex
- Quadratic
Classical bound: Lyapunov approach

\[ \mathbb{E} \left[ \| \theta_{k+1} - \theta^* \|^2 | \mathcal{F}_k \right] \leq \mathbb{E} \left[ \| \theta_k - \theta^* \|^2 \right] - 2\eta_k \langle F'(\theta_k), \theta_k - \theta^* \rangle \\
\quad + \eta_k^2 \| g_k(\theta_k) \|^2 \\
\leq \mathbb{E} \left[ \| \theta_k - \theta^* \|^2 \right] - 2\eta_k (1 - \eta_k L) \langle F'(\theta_k), \theta_k - \theta^* \rangle \\
\quad + \eta_k^2 \| g_k(\theta^*) \|^2 \\
\eta_k (F(\theta_k) - F(\theta^*)) \leq (1 - \eta_k \mu) \mathbb{E} \left[ \| \theta_k - \theta^* \|^2 \right] - \mathbb{E} \left[ \| \theta_{k+1} - \theta^* \|^2 | \mathcal{F}_k \right] \\
\quad + \eta_k^2 \| g_k(\theta^*) \|^2 \]
Classical bound: Lyapunov approach

\[ \mathbb{E} \left[ \| \theta_{k+1} - \theta^* \|^2 | \mathcal{F}_k \right] \leq \mathbb{E} \left[ \| \theta_k - \theta^* \|^2 \right] - 2\eta_k \langle F'(\theta_k), \theta_k - \theta^* \rangle \\
+ \eta_k^2 \| g_k(\theta_k) \|^2 \\
\leq \mathbb{E} \left[ \| \theta_k - \theta^* \|^2 \right] - 2\eta_k (1 - \eta_k L) \langle F'(\theta_k), \theta_k - \theta^* \rangle \\
+ \eta_k^2 \| g_k(\theta^*) \|^2 \\
\eta_k (F(\theta_k) - F(\theta^*)) \leq (1 - \eta_k \mu) \mathbb{E} \left[ \| \theta_k - \theta^* \|^2 \right] - \mathbb{E} \left[ \| \theta_{k+1} - \theta^* \|^2 | \mathcal{F}_k \right] \\
+ \eta_k^2 \| g_k(\theta^*) \|^2 \\
\]

Conclusion: with \( \eta_k = \frac{1}{\mu_k} \), telescopic sum + Jensen:

\[ \mathbb{E} \left[ F(\tilde{\theta}_k) - F(\theta^*) \right] \leq O(1/\mu k). \]
Trivial case: decaying step sizes are not that great!

Consider least squares: \( y_i = \theta^\top x_i + \varepsilon_i, \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \)
Consider least squares: $y_i = \theta^* \top x_i + \varepsilon_i$, $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Start with $\theta_0 = \theta^*$:

Then:

$$\bar{\theta}_k - \theta^* = \frac{1}{k} \sum_{i=1}^{k} M_i^k \eta_i^2 \varepsilon_i.$$

Even with large step size $\eta_i^2 \equiv \eta$, CLT is enough to control that!
Trivial case: decaying step sizes are not that great!

Consider least squares: \( y_i = \theta^* \top x_i + \varepsilon_i, \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \).

Start with \( \theta_0 = \theta^* \):

Then:

\[
\bar{\theta}_k - \theta^* = \frac{1}{k} \sum_{i=1}^{k} M_i^k \eta_i^2 \varepsilon_i.
\]

Even with large step size \( \eta_i^2 \equiv \eta \), CLT is enough to control that!

Tight control is much easier on the stochastic process \( \theta_k - \theta^* \) than through the “Lyapunov approach”.
Other proof: introduce decomposition


\[ \eta_k F''(\theta^*)(\theta_{k-1} - \theta^*) = \theta_{k-1} - \theta_k \]

\[ -\eta_k \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]

\[ + \eta_k \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]
Other proof: introduce decomposition


\[ \eta_k F''(\theta^*)(\theta_{k-1} - \theta^*) = \theta_{k-1} - \theta_k \]
\[-\eta_k \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]
\[+ \eta_k \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Thus, for \( \eta_k \equiv \eta \)

\[ F''(\theta^*)(\bar{\theta}_K - \theta^*) = \frac{\theta_K - \theta_0}{\eta K} - \frac{1}{K} \sum_{k=1}^{K} \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]
\[+ \frac{1}{K} \sum_{k=1}^{K} \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]
Other proof: introduce decomposition


\[ \eta_k F''(\theta^*)(\theta_{k-1} - \theta^*) = \theta_{k-1} - \theta_k \]

\[ - \eta_k \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]

\[ + \eta_k \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Thus, for \( \eta_k \equiv \eta \)

\[ F''(\theta^*)(\bar{\theta}_K - \theta^*) = \frac{\theta_K - \theta_0}{\eta K} - \frac{1}{K} \sum_{k=1}^{K} \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]

\[ + \frac{1}{K} \sum_{k=1}^{K} \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Initial condition - Noise - Non quadratic residual
Other proof: introduce decomposition


\[ \eta_k F''(\theta^*)(\theta_{k-1} - \theta^*) = \theta_{k-1} - \theta_k \]
\[ - \eta_k \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]
\[ + \eta_k \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Thus, for \( \eta_k \equiv \eta \)

\[ F''(\theta^*)(\bar{\theta}_K - \theta^*) = \frac{\theta_K - \theta_0}{\eta K} - \frac{1}{K} \sum_{k=1}^{K} \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]
\[ + \frac{1}{K} \sum_{k=1}^{K} \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Initial condition - Noise - Non quadratic residual

\( \leftrightarrow \) tight control of \( \| F''(\theta^*)(\bar{\theta}_K - \theta^*) \| \).
Other proof: introduce decomposition


\[ \eta_k F''(\theta^*)(\theta_{k-1} - \theta^*) = \theta_{k-1} - \theta_k \]

\[ -\eta_k \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]

\[ + \eta_k \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Thus, for \( \eta_k \equiv \eta \)

\[ F''(\theta^*) (\bar{\theta}_K - \theta^*) = \frac{\theta_K - \theta_0}{\eta K} - \frac{1}{K} \sum_{k=1}^{K} \left[ g_k(\theta_{k-1}) - F'(\theta_{k-1}) \right] \]

\[ + \frac{1}{K} \sum_{k=1}^{K} \left[ F'(\theta_{k-1}) - F''(\theta^*)(\theta_{k-1} - \theta^*) \right]. \]

Initial condition - Noise - Non quadratic residual

\[ \Leftrightarrow \] tight control of \( \| F''(\theta^*) (\bar{\theta}_K - \theta^*) \| \).

Correct control of the noise for smooth and strongly convex

All step sizes \( \eta_n = C n^{-\alpha} \) with \( \alpha \in (1/2, 1) \) lead to \( O(n^{-1}) \).

LMS algorithm: constant step-size \( \rightarrow \) statistical optimality.
Problem: dependence in $\mu$

Possible to recover convergence in function values:

$$F(\bar{\theta}_K) - F(\theta^*) \leq \frac{L}{2} \|\theta_K - \theta^*\|^2 \leq \frac{L}{2\mu^2} \|F''(\theta^*) (\bar{\theta}_K - \theta^*)\|^2$$

However:
▶ Ok for least squares regression (with some more work (Défossez and Bach, 2015; Dieuleveut et al., 2016; Jain et al., 2016)).
▶ Possible to recover tight convergence with self concordance (Bach 2013).
Possible to recover convergence in function values:

\[ F(\bar{\theta}_K) - F(\theta^*) \leq \frac{L}{2} \| \theta_K - \theta^* \|^2 \leq \frac{L}{2\mu^2} \| F''(\theta^*) (\bar{\theta}_K - \theta^*) \|^2 \]

However:

- Ok for least squares regression (with some more work (Défossez and Bach, 2015; Dieuleveut et al., 2016; Jain et al., 2016))
- Possible to recover tight convergence with self concordance (Bach 2013).
Synchronized distributed optimization

1. $P$ machines
2. $C$ the number of communication steps (C phases)
3. for $t \in [C]$, worker $p \in [P]$ performs $N^t$ local steps
Synchronized distributed optimization

1. $P$ machines
2. $C$ the number of communication steps (C phases)
3. for $t \in [C]$, worker $p \in [P]$ performs $N^t$ local steps

For any $p \in [P]$, $t \in [C]$, $k \in [N^t]$:

- $\theta_{p,k}^t$ the model proposed by worker $p$, at phase $t$, after $k$ local iterations.
- $\theta_{p,0}^1 = \theta_0$.
- $\theta_{p,k}^t = \theta_{p,k-1}^t - \eta_k^t g_{p,k}^t(\theta_{p,k-1}^t)$.
Link with classical algorithms.

<table>
<thead>
<tr>
<th>Algo.</th>
<th>Work.</th>
<th>Com.</th>
<th>Phases</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>$P$</td>
<td>$C$</td>
<td>$(N^1 \ldots N^C)$</td>
<td>$P \sum_{t=1}^{C} N^t$</td>
</tr>
<tr>
<td>Serial</td>
<td>1</td>
<td>-</td>
<td>$(N)$</td>
<td>$P \frac{N}{N}$</td>
</tr>
<tr>
<td>P-MBA</td>
<td>$P$</td>
<td>$C$</td>
<td>$(1, \ldots, 1)$</td>
<td>$PC$</td>
</tr>
<tr>
<td>OSA</td>
<td>$P$</td>
<td>1</td>
<td>$(N^1)$</td>
<td>$N^1P$</td>
</tr>
</tbody>
</table>
Link with classical algorithms.

<table>
<thead>
<tr>
<th>Algo.</th>
<th>Work.</th>
<th>Com.</th>
<th>Phases</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>$P$</td>
<td>$C$</td>
<td>$(N^1 \ldots N^C)$</td>
<td>$P \sum_{t=1}^{C} N^t$</td>
</tr>
<tr>
<td>Serial</td>
<td>1</td>
<td>-</td>
<td>$(N)$</td>
<td>$PC$</td>
</tr>
<tr>
<td>P-MBA</td>
<td>$P$</td>
<td>$C$</td>
<td>$(1, \ldots, 1)$</td>
<td>$N^1 P$</td>
</tr>
<tr>
<td>OSA</td>
<td>$P$</td>
<td>1</td>
<td>$(N^1)$</td>
<td>$N^1 P$</td>
</tr>
</tbody>
</table>

One Shot Averaging – Mini-Batch Averaging – Local SGD
Aggregation steps: $\hat{\theta}^t = \frac{1}{P} \sum_{p=1}^{P} \theta_{p,N_t}^t$.

At phase $t+1$, every worker $p \in [P]$ restarts from the averaged model: $\theta_{p,0}^{t+1} := \hat{\theta}^t$.

Goal: Risk of the Polyak-Ruppert averaged iterate:

$$\overline{\theta}^C = \frac{1}{P \sum_{t=1}^{C} N_t} \sum_{t=1}^{C} \sum_{p=1}^{P} \sum_{k=1}^{N_t} \theta_{p,k}^t,$$
Assumptions

**A3** [Oracle on the gradient] Filtration $(\mathcal{H}_k^t)_{(t,k)\in[C] \times [N^t]}$ such that for any $(t, k) \in [C] \times [N^t]$ and $\theta \in \mathbb{R}^d$, $g_{p,k+1}^t(\theta)$ is a $\mathcal{H}_{k+1}^t$-measurable random variable and $\mathbb{E}\left[ g_{p,k+1}^t(\theta) | \mathcal{H}_k^t \right] = F'(\theta)$.

**A4** [Uniformly bounded variance]
$\mathbb{E}[\|g_{p,k}^t(\theta_{p,k}^t) - F'(\theta_{p,k}^t)\|^2] \leq \sigma_\infty^2$.

**A5** [Cocoercivity of the random gradients] For any $t \in [C]$, $k \in [N^t]$, $p \in [P]$, $g_{p,k}^t$ is almost surely $L$-co-coercive

**A6** [Finite variance at the optimal point] There exists $\sigma \geq 0$, such that for any $t \in [C]$, $k \in [N^t]$, $p \in [P]$, $\mathbb{E}[\|g_{p,k}^t(\theta^*)\|^4] \leq \sigma^4$.

We assume **A4** OR **A5 + A6**
Error decomposition

\[ \eta_k^t F''(\theta^*)(\theta_{p,k-1}^t - \theta^*) = \theta_{p,k-1}^t - \theta_{p,k}^t - \eta_k^t \left[ g_{p,k}^t(\theta_{p,k-1}^t) - F'(\theta_{p,k-1}^t) \right] + \eta_k^t \left[ F'(\theta_{p,k-1}^t) - F''(\theta^*)(\theta_{p,k-1}^t - \theta^*) \right]. \]

Thus:

\[ F''(\theta^*) \left( \frac{\bar{\theta}^C}{\theta} - \theta^* \right) = \frac{1}{P \sum_{t=1}^C N^t} \sum_{t=1}^C \sum_{p=1}^P \sum_{k=1}^{N^t} \left( \frac{\theta_{p,k-1}^t - \theta_{p,k}^t}{\eta_k^t} \right) - \left[ g_{p,k}^t(\theta_{p,k-1}^t) - F'(\theta_{p,k-1}^t) \right] + \left[ F'(\theta_{p,k-1}^t) - F''(\theta^*)(\theta_{p,k-1}^t - \theta^*) \right]. \]
Error decomposition

\[ \eta_k^t F''(\theta^*)(\theta_{p,k-1}^t - \theta^*) = \theta_{p,k-1}^t - \theta_{p,k}^t \]

\[- \eta_k^t \left[ g_{p,k}^t(\theta_{p,k-1}^t) - F'(\theta_{p,k-1}^t) \right] \]

\[+ \eta_k^t \left[ F'(\theta_{p,k-1}^t) - F''(\theta^*)(\theta_{p,k-1}^t - \theta^*) \right]. \]

Thus:

\[ F''(\theta^*) \left( \bar{\theta}^C - \theta^* \right) = \frac{1}{P \sum_{t=1}^{C} \sum_{p=1}^{P} \sum_{k=1}^{N^t} \left( \frac{\theta_{p,k-1}^t - \theta_{p,k}^t}{\eta_k^t} \right)} \]

\[- \left[ g_{p,k}^t(\theta_{p,k-1}^t) - F'(\theta_{p,k-1}^t) \right] \]

\[+ \left[ F'(\theta_{p,k-1}^t) - F''(\theta^*)(\theta_{p,k-1}^t - \theta^*) \right]. \]

Noise: Additive $+$ (Multiplicative $\propto ||\theta_{p,k}^t - \theta^*||^2$)

Residual: $\propto ||\theta_{p,k}^t - \theta^*||^2$
Assume $A1,2,3,5,6$, and $\eta_k \equiv \eta$ for any $(t, k) \in [C] \times [N^t]$.

**Proposition (Mini-batch Averaging)**

For any $t \in [C],$

$$
E \left[ \|\hat{\theta}^t - \theta^*\|^2 \right] \leq (1 - \eta \mu)^t \|\theta_0 - \theta^*\|^2 + \frac{2\sigma^2 \eta (1 - (1 - \eta \mu)^t)}{P},
$$

$$
E \left[ \|\overline{\theta}_C - \theta^*\|^2_{F''(\theta^*)} \right] \lesssim \frac{\|\theta_0 - \theta^*\|^2}{\eta^2 C^2} Q_{bias} + \frac{\sigma^2}{T} \left(1 + \frac{Q_{1,\text{var}(C)}}{P} + \frac{Q_{2,\text{var}(C)}}{P^2}\right).
$$
Results MBA - OSA

Assume A1,2,3,5,6, and $\eta^t_k \equiv \eta$ for any $(t, k) \in [C] \times [N^t]$.

Proposition (Mini-batch Averaging)

For any $t \in [C]$,

$$
\mathbb{E} \left[ \left\| \hat{\theta}^t - \theta^* \right\|^2 \right] \leq (1 - \eta \mu)^t \left\| \theta_0 - \theta^* \right\|^2 + \frac{2\sigma^2 \eta \left( 1 - (1 - \eta \mu)^t \right)}{P},
$$

$$
\mathbb{E} \left[ \left\| \overline{\theta}^C - \theta^* \right\|^2_{F''(\theta^*)} \right] \preceq \frac{\left\| \theta_0 - \theta^* \right\|^2}{\eta^2 C^2} Q_{bias} + \frac{\sigma^2}{T} \left( 1 + \frac{Q_{1, var}(C)}{P} + \frac{Q_{2, var}(C)}{P^2} \right).
$$

Proposition (One-shot Averaging)

For any $p \in [P]$, $t = 1$, $k \in [N]$,

$$
\mathbb{E} \left[ \left\| \theta^1_{p,k} - \theta^* \right\|^2 \right] \leq (1 - \eta \mu)^k \left\| \theta_0 - \theta^* \right\|^2 + 2\sigma^2 \eta \left( 1 - (1 - \eta \mu)^k \right),
$$

$$
\mathbb{E} \left[ \left\| \overline{\theta}^C - \theta^* \right\|^2_{F''(\theta^*)} \right] \preceq \frac{\left\| \theta_0 - \theta^* \right\|^2}{\eta^2 N^2} Q_{bias} + \frac{\sigma^2}{T} \left( 1 + Q_{1, var}(N) + Q_{2, var}(N) \right).
$$

Total number of gradients processed is $T = PC$ resp $T = PN$. 
With

\[ Q_{\text{bias}} = 1 + \frac{M^2 \eta}{\mu} \left\| \theta^0 - \theta^* \right\|^2 + \frac{L^2 \eta}{\mu P}, \]

\[ Q_{1,\text{var}}(X) = \frac{L^2 \eta}{\mu} + \frac{P}{X \eta \mu}, \quad Q_{2,\text{var}}(X) = \frac{M^2 X P \eta^2 \sigma^2}{\mu^2}. \]
With

\[ Q_{\text{bias}} = 1 + \frac{M^2 \eta}{\mu} \left\| \theta^0 - \theta^* \right\|^2 + \frac{L^2 \eta}{\mu P}, \]

\[ Q_{1,\text{var}}(X) = \frac{L^2 \eta}{\mu} + \frac{P}{X \eta \mu}, \quad Q_{2,\text{var}}(X) = \frac{M^2 XP \eta^2 \sigma^2}{\mu^2}. \]

- Asymptotically equivalent for \( P \) constant.
- Non asymptotic result (vs Godichon and Saadane (2017)).
- Proposition 1 corrects Bach 2011, with Needel 2014 remark (see also Dieuleveut Durmus 2017).
- “the noise is the noise and SGD doesn’t care” (for asynchronous SGD, (Duchi et al., 2015))
- Extension to the on-line setting possible
Assume Q1, A3, A4. For \( p \in [P] \), \( t \in [C] \), \( k \in [N^t] \),

\[
\mathbb{E} \left[ \left\| \hat{\theta}^{t-1} - \theta^* \right\|^2 \right] \leq (1 - \eta \mu)^{N^t_1 - 1} \left\| \theta_0 - \theta^* \right\|^2 + \frac{\sigma^2_\infty \eta}{P} \frac{1 - (1 - \eta \mu)^{N^t_1 - 1}}{\mu}
\]

\[
\mathbb{E} \left[ \left\| \theta_{p,k}^t - \theta^* \right\|^2 \right] \leq (1 - \eta \mu)^{N^t_1 - 1 + k} \left\| \theta_0 - \theta^* \right\|^2 + \sigma^2_\infty \eta \left( \frac{1 - (1 - \eta \mu)^{N^t_1 - 1}}{P \mu} \text{ long term reduced variance} + \frac{1 - (1 - \eta \mu)^k}{\mu} \text{ local iteration variance} \right).
\]

**Corollary:** If for all \( t \in [C] \), \( N^t \leq \frac{1}{\mu \eta P} \), then the second order moment of \( \theta_{p,k}^t \) admits the same upper bound as the mini-batch iterate \( \hat{\theta}_{MB}^{N^t_1 - 1 + k} \) up to a factor of 2. As a consequence, **Local-SGD performs optimally**.
Bridging the gap: convergence of Local-SGD: simple case

Assume Q1, A3, A4. For $p \in [P]$, $t \in [C]$, $k \in [N^t]$, 

$$
\mathbb{E}\left[\|\hat{\theta}_{t}^{t-1} - \theta^*\|^2\right] \leq (1 - \eta \mu)^{N^t_1-1} \|\theta_0 - \theta^*\|^2 + \frac{\sigma^2 \eta}{P} \frac{1 - (1 - \eta \mu)^{N^t_1-1}}{\mu} 
$$

$$
\mathbb{E}\left[\|\theta_{p,k}^{t} - \theta^*\|^2\right] \leq (1 - \eta \mu)^{N^t_1-1+k} \|\theta_0 - \theta^*\|^2 + \sigma^2 \eta \left(\frac{1 - (1 - \eta \mu)^{N^t_1-1}}{P \mu} + \frac{1 - (1 - \eta \mu)^{k}}{\mu}\right).
$$

**Corollary:** If for all $t \in [C]$, $N^t \leq \frac{1}{\mu \eta P}$, then the second order moment of $\theta_{p,k}^t$ admits the same upper bound as the mini-batch iterate $\hat{\theta}_{MB}^{N^t_1-1+k}$ up to a factor of 2. As a consequence, **Local-SGD performs optimally.**
Example

With constant number of local steps $N^t = N$, and learning rate $\eta = \frac{c}{\sqrt{NC}}$ in order to obtain an optimal $O(\frac{\sigma^2}{T})$ parallel convergence rate, local-SGD can communicate $O(\frac{\sqrt{NC}}{P\mu})$ times less as compared to mini-batch averaging.
Quadratic + additive noise $\leftrightarrow$ too simple and un-realistic

- Least square regression: quadratic $+$ multiplicative noise (Q1, A3, A5, A6)
- Logistic regression: non quadratic $+$ uniformly bounded variance (A1, A2, A3, A4)

Key lemmas: control how the restart point of each phase differs from its mini-batch equivalent.
Quadratic + additive noise ⇔ too simple and un-realistic

- Least square regression: quadratic + multiplicative noise (Q1, A3, A5, A6)
- Logistic regression: non quadratic + uniformly bounded variance (A1, A2, A3, A4)

Key lemmas: control how the restart point of each phase differs from its mini-batch equivalent.

**Theorem**
Under either of the following sets of assumptions, the convergence of the Polyak Ruppert iterate $\bar{\theta}$ is as good as in the mini-batch case, up to a constant:

1. Assume Q1, A3, A5, A6, and for any $t \in [C]$, $N_t \leq \frac{1}{\mu \eta P}$ and $\mu \eta^2 N_1^t = O(1)$.
Quadratic + additive noise ↔ too simple and un-realistic

- Least square regression: quadratic + multiplicative noise (Q1, A3, A5, A6)
- Logistic regression: non quadratic + uniformly bounded variance (A1, A2, A3, A4)

Key lemmas: control how the restart point of each phase differs from its mini-batch equivalent.

**Theorem**

Under either of the following sets of assumptions, the convergence of the Polyak Ruppert iterate $\overline{\theta}$ is as good as in the mini-batch case, up to a constant:

1. Assume Q1, A3, A5, A6, and for any $t \in [C]$, $N_t \leq \frac{1}{\mu \eta P}$ and $\mu \eta^2 N_1^t = O(1)$.

2. Assume A1, A2, A3, A4, and for any $t \in [C]$, $N_t \leq \inf \left( \frac{1}{\eta PM \mathbb{E} [\|\hat{\theta}^t - \theta^*\|]}, \frac{1}{\mu \eta P} \right)$. 
Conclusion

- Non asymptotic analysis of Local-SGD
- With “large” step sizes.
- Better understanding of communication trade-offs → lower bounds on communication frequency
- Similar results for the on-line case (a bit faster, and much more painful for the eyes).
Conclusion

Non asymptotic analysis of Local-SGD

With “large” step sizes.

Better understanding of communication trade-offs → lower bounds on communication frequency

Similar results for the on-line case (a bit faster, and much more painful for the eyes).

Directions:

Improve to optimal rates in terms of \( \mu \) with self concordance

Proving that those bounds are tight (dangerous to compare upper bounds!!)


