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.**

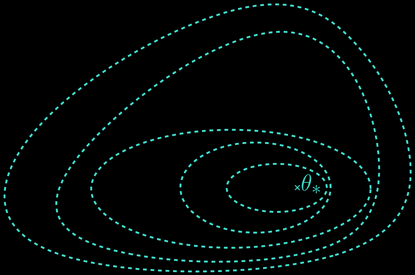
Stochastic Gradient Descent

- **Goal:**

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

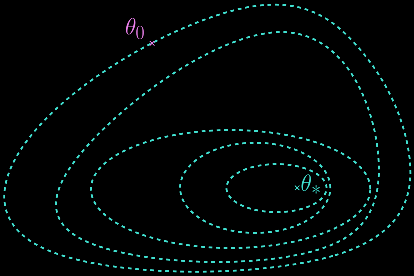
  given unbiased gradient estimates g_n

- $\theta^\star := \operatorname{argmin}_{\mathbb{R}^d} F(\theta)$.



Stochastic Gradient Descent

- **Goal:**
$$\min_{\theta \in \mathbb{R}^d} F(\theta)$$
given unbiased gradient estimates g_n
- $\theta^\star := \mathrm{argmin}_{\mathbb{R}^d} F(\theta)$.

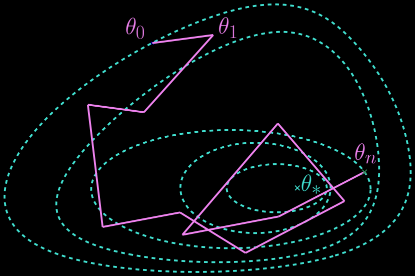


- Key algorithm: Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951):

$$\boxed{\theta_k = \theta_{k-1} - \eta_k \, g_k(\theta_{k-1})}$$

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

Stochastic Gradient Descent



- **Goal:**

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

  given unbiased gradient estimates g_n
- $\theta^\star := \operatorname{argmin}_{\mathbb{R}^d} F(\theta)$.
- **Key algorithm: Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951):**

$$\boxed{\theta_k = \theta_{k-1} - \eta_k \, g_k(\theta_{k-1})}$$

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

Supervised Machine Learning

- We define the risk (generalization error) as

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

- 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

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

- 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:

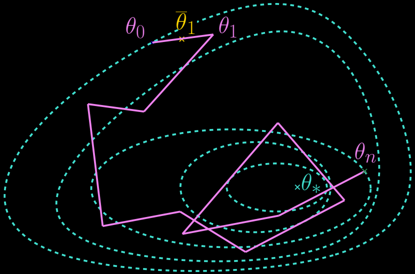
$$\min_{\theta\in\mathbb{R}^d} \quad \frac{1}{2n}\sum_{i=1}^{n} \left(y_i - \langle\theta, \Phi(x_i)\rangle\right)^2 \quad + \quad \mu\Omega(\theta),$$

- and logistic regression:

$$\min_{\theta\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-y_i\langle\theta, \Phi(x_i)\rangle)\right) \quad + \quad \mu\Omega(\theta).$$

Polyak Ruppert averaging

Polyak Ruppert averaging

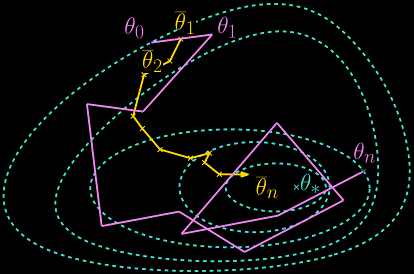


Introduced by Polyak and Juditsky (1992) and Ruppert (1988):

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

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

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

Assumptions

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

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: $\sup_{\theta \in \mathbb{R}^d} \left|\left|\left| F^{(2)}(\theta) \right|\right|\right| < L$, and $\sup_{\theta \in \mathbb{R}^d} \left|\left|\left| F^{(3)}(\theta) \right|\right|\right| < M$. Especially F is L-smooth.

Assumptions

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

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.

Or:

Q1 [Quadratic function] There exists a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, such that the function F is the quadratic function $\theta \mapsto \|\boldsymbol{\Sigma}^{1/2}(\theta - \theta^\star)\|^2/2$,

Which step size would you use?

Smooth functions.

	$\eta_k \equiv \eta_0$	$\eta_k = 1/\sqrt{k}$	$\eta_k = 1/(\mu k)$
Convex			
Strongly Convex			
Quadratic			

Classical bound: Lyapunov approach

$$
\begin{aligned}
\mathbb{E}\left[||\theta_{k+1} - \theta^\star||^2 | \mathcal{F}_k\right] &\leq \mathbb{E}\left[||\theta_k - \theta^\star||^2\right] - 2\eta_k \left\langle F'(\theta_k), \theta_k - \theta^\star \right\rangle \\
&\quad + \eta_k^2 ||g_k(\theta_k)||^2 \\
&\leq \mathbb{E}\left[||\theta_k - \theta^\star||^2\right] - 2\eta_k(1 - \eta_k L) \left\langle F'(\theta_k), \theta_k - \theta^\star \right\rangle \\
&\quad + \eta_k^2 ||g_k(\theta^\star)||^2
\end{aligned}
$$

$$
\begin{aligned}
\eta_k(F(\theta_k) - F(\theta^\star)) \leq (1 - \eta_k \mu)\mathbb{E}\left[||\theta_k - \theta^\star||^2\right] - \mathbb{E}\left[||\theta_{k+1} - \theta^\star||^2 | \mathcal{F}_k\right] \\
+ \eta_k^2 ||g_k(\theta^\star)||^2
\end{aligned}
$$

Classical bound: Lyapunov approach

$$
\begin{aligned}
\mathbb{E}\left[||\theta_{k+1}-\theta^\star||^2|\mathcal{F}_k\right] &\leq \mathbb{E}\left[||\theta_k-\theta^\star||^2\right] - 2\eta_k\left\langle F'(\theta_k), \theta_k-\theta^\star\right\rangle \\
&\quad + \eta_k^2||g_k(\theta_k)||^2 \\
&\leq \mathbb{E}\left[||\theta_k-\theta^\star||^2\right] - 2\eta_k(1-\eta_k L)\left\langle F'(\theta_k), \theta_k-\theta^\star\right\rangle \\
&\quad + \eta_k^2||g_k(\theta^\star)||^2
\end{aligned}
$$

$$
\begin{aligned}
\eta_k(F(\theta_k)-F(\theta^\star)) &\leq (1-\eta_k\mu)\mathbb{E}\left[||\theta_k-\theta^\star||^2\right] - \mathbb{E}\left[||\theta_{k+1}-\theta^\star||^2|\mathcal{F}_k\right] \\
&\quad + \eta_k^2||g_k(\theta^\star)||^2
\end{aligned}
$$

Conclusion: with $\eta_k = \frac{1}{\mu k}$, telescopic sum + Jensen:

$$
\mathbb{E}\left[F(\bar{\theta}_k) - F(\theta^\star)\right] \leq O(1/\mu k).
$$

Trivial case: decaying step sizes are not that great !

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

Trivial case: decaying step sizes are not that great !

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

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

Then:

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

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

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

Then:

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

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 $\theta_k - \theta^\star$ than through the "Lyapunov approach".

Other proof: introduce decomposition

Original proof of averaging in Polyak and Juditsky (1992).

$$\begin{aligned}\eta_k F''(\theta^\star)(\theta_{k-1}-\theta^\star) = {} & \theta_{k-1}-\theta_k \\ & -\eta_k\left[g_k(\theta_{k-1}) - F'(\theta_{k-1})\right] \\ & +\eta_k\left[F'(\theta_{k-1}) - F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Other proof: introduce decomposition

Original proof of averaging in Polyak and Juditsky (1992).

$$\begin{aligned}\eta_k F''(\theta^\star)(\theta_{k-1} - \theta^\star) = {} & \theta_{k-1} - \theta_k \\ & -\eta_k \left[g_k(\theta_{k-1}) - F'(\theta_{k-1})\right] \\ & + \eta_k \left[F'(\theta_{k-1}) - F''(\theta^\star)(\theta_{k-1} - \theta^\star)\right].\end{aligned}$$

Thus, for $\eta_k \equiv \eta$

$$\begin{aligned}F''(\theta^\star)\left(\bar\theta_K - \theta^\star\right) = {} & \frac{\theta_K - \theta_0}{\eta K} - \frac{1}{K}\sum_{k=1}^{K}\left[g_k(\theta_{k-1}) - F'(\theta_{k-1})\right] \\ & + \frac{1}{K}\sum_{k=1}^{K}\left[F'(\theta_{k-1}) - F''(\theta^\star)(\theta_{k-1} - \theta^\star)\right].\end{aligned}$$

Other proof: introduce decomposition

Original proof of averaging in Polyak and Juditsky (1992).

$$\begin{aligned}\eta_k F''(\theta^\star)(\theta_{k-1}-\theta^\star) &= \theta_{k-1}-\theta_k \\ &\quad -\eta_k\left[g_k(\theta_{k-1})-F'(\theta_{k-1})\right] \\ &\quad +\eta_k\left[F'(\theta_{k-1})-F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Thus, for $\eta_k \equiv \eta$

$$\begin{aligned}F''(\theta^\star)\left(\bar\theta_K-\theta^\star\right) &= \frac{\theta_K-\theta_0}{\eta K}-\frac{1}{K}\sum_{k=1}^{K}\left[g_k(\theta_{k-1})-F'(\theta_{k-1})\right] \\ &\quad +\frac{1}{K}\sum_{k=1}^{K}\left[F'(\theta_{k-1})-F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Initial condition - Noise - Non quadratic residual

Other proof: introduce decomposition

Original proof of averaging in Polyak and Juditsky (1992).

$$\begin{aligned}\eta_k F''(\theta^\star)(\theta_{k-1}-\theta^\star) = {}& \theta_{k-1}-\theta_k \\ & -\eta_k\left[g_k(\theta_{k-1}) - F'(\theta_{k-1})\right] \\ & +\eta_k\left[F'(\theta_{k-1}) - F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Thus, for $\eta_k \equiv \eta$

$$\begin{aligned}F''(\theta^\star)\left(\bar\theta_K-\theta^\star\right) = {}& \frac{\theta_K-\theta_0}{\eta K} - \frac{1}{K}\sum_{k=1}^{K}\left[g_k(\theta_{k-1}) - F'(\theta_{k-1})\right] \\ & + \frac{1}{K}\sum_{k=1}^{K}\left[F'(\theta_{k-1}) - F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Initial condition - Noise - Non quadratic residual

$\looparrowright$ tight control of $||F''(\theta^\star)\left(\bar\theta_K-\theta^\star\right)||$.

Other proof: introduce decomposition

Original proof of averaging in Polyak and Juditsky (1992).

$$\begin{aligned}\eta_k F''(\theta^\star)(\theta_{k-1}-\theta^\star) &= \theta_{k-1}-\theta_k \\ &\quad -\eta_k\left[g_k(\theta_{k-1})-F'(\theta_{k-1})\right] \\ &\quad +\eta_k\left[F'(\theta_{k-1})-F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Thus, for $\eta_k \equiv \eta$

$$\begin{aligned}F''(\theta^\star)\left(\bar\theta_K-\theta^\star\right) &= \frac{\theta_K-\theta_0}{\eta K}-\frac{1}{K}\sum_{k=1}^{K}\left[g_k(\theta_{k-1})-F'(\theta_{k-1})\right] \\ &\quad +\frac{1}{K}\sum_{k=1}^{K}\left[F'(\theta_{k-1})-F''(\theta^\star)(\theta_{k-1}-\theta^\star)\right].\end{aligned}$$

Initial condition - Noise - Non quadratic residual

$\looparrowright$ tight control of $\|F''(\theta^\star)\left(\bar\theta_K-\theta^\star\right)\|$.

Correct control of the noise for smooth and strongly convex

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

LMS algorithm: constant step-size $\to$ statistical optimality.

Problem: dependence in μ

Possible to recover convergence in function values:

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

Problem: dependence in μ

Possible to recover convergence in function values:

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

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

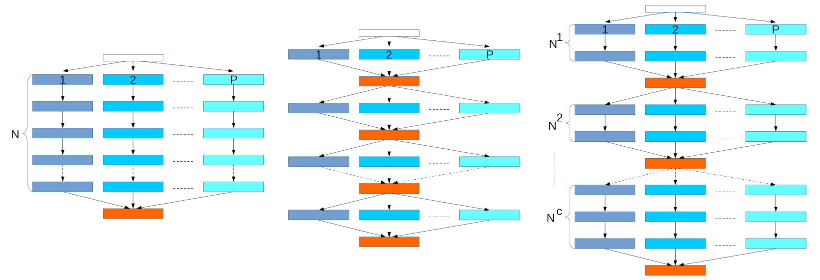
- $\theta^t_{p,k}$ the model proposed by worker p, at phase t, after k local iterations.
- $\theta^1_{p,0} = \theta_0$.
- $$\boxed{\theta^t_{p,k} = \theta^t_{p,k-1} - \eta^t_k g^t_{p,k}(\theta^t_{p,k-1}).}$$

Link with classical algorithms.

Algo.	Work.	Com.	Phases	T
Local	P	C	$(N^1 \ldots N^C)$	$P\sum_{t=1}^{C} N^t$
Serial	1	-	(N)	N
P-MBA	P	C	$(1, \ldots, 1)$	PC
OSA	P	1	(N^1)	$N^1 P$

Link with classical algorithms.

Algo.	Work.	Com.	Phases	T
Local	P	C	$(N^1 \dots N^C)$	$P\sum_{t=1}^{C} N^t$
Serial	1	-	(N)	N
P-MBA	P	C	$(1, \dots, 1)$	PC
OSA	P	1	(N^1)	$N^1 P$



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:

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

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 $t\in[C], k\in[N^t], p\in[P]$, $\mathbb{E}[\|g_{p,k}^t(\theta^\star)\|^4]\leq\sigma^4$.

We assume A4 OR A5 + A6

Error decomposition

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

Thus:

$$\begin{aligned}F''(\theta^\star)\left(\overline{\overline{\theta}}^C - \theta^\star\right) = \frac{1}{P\sum_{t=1}^C N^t} \sum_{t=1}^C \sum_{p=1}^P \sum_{k=1}^{N^t} \Bigg(\frac{\theta_{p,k-1}^t - \theta_{p,k}^t}{\eta_k^t} \\ - \left[g_{p,k}^t(\theta_{p,k-1}^t) - F'(\theta_{p,k-1}^t) \right] \\ + \left[F'(\theta_{p,k-1}^t) - F''(\theta^\star)(\theta_{p,k-1}^t - \theta^\star) \right] \Bigg).\end{aligned}$$

Error decomposition

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

Thus:

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

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

Results MBA - OSA

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

Proposition (Mini-batch Averaging)

For any $t \in [C]$,

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

$$\mathbb{E}\left[\left\|\overline{\overline{\theta}}^C - \theta^\star\right\|^2_{F''(\theta^\star)}\right] \precsim \frac{\left\|\theta^0 - \theta^\star\right\|^2}{\eta^2 C^2} Q_{bias} + \frac{\sigma^2}{T}\left(1 + \frac{Q_{1,var}(C)}{P} + \frac{Q_{2,var}(C)}{P^2}\right).$$

Results MBA - OSA

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

Proposition (Mini-batch Averaging)

For any $t \in [C]$,

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

$$\mathbb{E}\left[\left\|\overline{\overline{\theta}}^C - \theta^\star\right\|^2_{F''(\theta^\star)}\right] \precsim \frac{\left\|\theta^0 - \theta^\star\right\|^2}{\eta^2 C^2} Q_{bias} + \frac{\sigma^2}{T}\left(1 + \frac{Q_{1,var}(C)}{P} + \frac{Q_{2,var}(C)}{P^2}\right).$$

Proposition (One-shot Averaging)

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

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

$$\mathbb{E}\left[\left\|(\overline{\overline{\theta}}^C - \theta^\star)\right\|^2_{F''(\theta^\star)}\right] \precsim \frac{\left\|\theta^0 - \theta^\star\right\|^2}{\eta^2 N^2} Q_{bias} + \frac{\sigma^2}{T}\left(1 + Q_{1,var}(N) + Q_{2,var}(N)\right)$$

With

$$Q_{bias} = 1 + \frac{M^2\eta}{\mu}\left\|\theta^0 - \theta^\star\right\|^2 + \frac{L^2\eta}{\mu P},$$

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

With

$$Q_{bias} = 1 + \frac{M^2\eta}{\mu}\left\|\theta^0 - \theta^\star\right\|^2 + \frac{L^2\eta}{\mu P},$$

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

- 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

Bridging the gap: convergence of Local-SGD: simple case

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

$$\mathbb{E}\left[\left\|\hat{\theta}^{t-1} - \theta^\star\right\|^2\right] \leq (1-\eta\mu)^{N_1^{t-1}} \|\theta_0 - \theta^\star\|^2 + \frac{\sigma_\infty^2 \eta}{P} \frac{1-(1-\eta\mu)^{N_1^{t-1}}}{\mu}$$

$$\mathbb{E}\left[\left\|\theta_{p,k}^t - \theta^\star\right\|^2\right] \leq (1-\eta\mu)^{N_1^{t-1}+k} \|\theta_0 - \theta^\star\|^2$$
$$+ \sigma_\infty^2 \eta \left(\underbrace{\frac{1-(1-\eta\mu)^{N_1^{t-1}}}{P\mu}}_{\text{long term reduced variance}} + \underbrace{\frac{1-(1-\eta\mu)^k}{\mu}}_{\text{local iteration variance}} \right).$$

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 $\hat{\theta}_{MB}^{N_1^{t-1}+k}$ 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 $p \in [P], t \in [C], k \in [N^t]$,

$$\mathbb{E}\left[\left\|\hat{\theta}^{t-1} - \theta^\star\right\|^2\right] \leq (1-\eta\mu)^{N_1^{t-1}} \|\theta_0 - \theta^\star\|^2 + \frac{\sigma_\infty^2 \eta}{P} \frac{1-(1-\eta\mu)^{N_1^{t-1}}}{\mu}$$

$$\mathbb{E}\left[\left\|\theta_{p,k}^t - \theta^\star\right\|^2\right] \leq (1-\eta\mu)^{N_1^{t-1}+k} \|\theta_0 - \theta^\star\|^2 + \sigma_\infty^2 \eta \left(\underbrace{\frac{1-(1-\eta\mu)^{N_1^{t-1}}}{P\mu}}_{\text{long term reduced variance}} + \underbrace{\frac{1-(1-\eta\mu)^k}{\mu}}_{\text{local iteration variance}} \right).$$

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 $\hat{\theta}_{MB}^{N_1^{t-1}+k}$ 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 $\leftrightarrow$ 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.

Theorem

Under either of the following sets of assumptions, the convergence of the Polyak Ruppert iterate $\overline{\overline{\theta}}^{C}$ 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 $\leftrightarrow$ 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.

Theorem

Under either of the following sets of assumptions, the convergence of the Polyak Ruppert iterate $\overline{\overline{\theta}}^C$ 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 $t \in [C]$, $N^t \leq \inf\left(\frac{1}{\eta P M \mathbb{E}\left[\left\|\hat{\theta}^t - \theta^\star\right\|\right]}, \frac{1}{\mu\eta P}\right)$.

Conclusion

Conclusion

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

Conclusion

Conclusion

- Non asymptotic analysis of Local-SGD
- With "large" step sizes.
- better understanding of communication trade-offs $\rightarrow$ 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 μ with self concordance
- Proving that those bounds are tight (dangerous to compare upper bounds!!)

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. Ann. Statist., 40(5):2452–2482.

Bach, F. and Moulines, E. (2011). Non-asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS'11, pages 451–459, USA. Curran Associates Inc.

Défossez, A. and Bach, F. (2015). Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS).

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. (2012). Optimal distributed online prediction using mini-batches. Journal of Machine Learning Research, 13(Jan):165–202.

Dieuleveut, A., Flammarion, N., and Bach, F. (2016). Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. ArXiv e-prints.

Duchi, J. C., Chaturapruek, S., and Ré, C. (2015). Asynchronous stochastic convex optimization. ArXiv e-prints.

Godichon, A. B. and Saadane, S. (2017). On the rates of convergence of Parallelized Averaged Stochastic Gradient Algorithms. ArXiv e-prints.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. (2016). Parallelizing Stochastic Approximation Through Mini-Batching and Tail-Averaging. ArXiv e-prints.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an $O(1/t)$ rate for the stochastic projected subgradient method. ArXiv e-prints 1212.2002.

Li, M., Zhang, T., Chen, Y., and Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 661–670. ACM.

Lin, T., Stich, S. U., and Jaggi, M. (2018). Don't Use Large Mini-Batches, Use Local SGD. ArXiv e-prints.

Nemirovsky, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. SIAM J. Control Optim., 30(4):838–855.

Rakhlin, A., Shamir, O., Sridharan, K., et al. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In ICML. Citeseer.

Robbins, H. and Monro, S. (1951). A stochastic approxiation method. The Annals of mathematical Statistics, 22(3):400–407.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Stich, S. U. (2018). Local SGD Converges Fast and Communicates Little. ArXiv e-prints.

Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. (2013). Mini-batch primal and dual methods for svms. In Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28, pages III–1022. JMLR. org.

Zhang, J., De Sa, C., Mitliagkas, I., and Ré, C. (2016). Parallel SGD: When does averaging help? ArXiv e-prints.

Zhang, Y., Wainwright, M. J., and Duchi, J. C. (2012). Communication-efficient algorithms for statistical optimization. In Advances in Neural Information Processing Systems, pages 1502–1510.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. J. (2010). Parallelized stochastic gradient descent. In Advances in neural information processing systems, pages 2595–2603.