Statistical Optimality of Stochastic Gradient Descent through Multiple Passes

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with Loucas Pillaud-Vivien and Alessandro Rudi

Toulouse - September 2018
Two-minute summary

• **Stochastic gradient descent for large-scale machine learning**
  - Processes observations one by one
Two-minute summary

- **Stochastic gradient descent for large-scale machine learning**
  - Processes observations one by one

- **Theory:** Single pass SGD is optimal

- **Practice:** Multiple pass SGD always works better
Two-minute summary

- **Stochastic gradient descent for large-scale machine learning**
  - Processes observations one by one

- **Theory**: Single pass SGD is optimal
  - Only for “easy” problems

- **Practice**: Multiple pass SGD always works better
  - Provable for “hard” problems
  - Quantification of required number of passes
  - Optimal statistical performance
  - Source and capacity conditions from kernel methods
Least-squares regression in finite dimension

• **Data:** $n$ observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}, i = 1, \ldots, n$, i.i.d.

• Prediction as **linear** functions $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathcal{H} = \mathbb{R}^d$
Least-squares regression in finite dimension

- **Data**: \( n \) observations \((x_i, y_i) \in \mathcal{X} \times \mathbb{R}, i = 1, \ldots, n, \text{ i.i.d.}\)

- Prediction as **linear** functions \( \langle \theta, \Phi(x) \rangle \) of features \( \Phi(x) \in \mathcal{H} = \mathbb{R}^d \)
  - Optimal prediction \( \theta_* \in \mathcal{H} \) minimizing \( F(\theta) = \mathbb{E}(y - \langle \theta, \Phi(x) \rangle)^2 \)
Least-squares regression in finite dimension

- **Data**: $n$ observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \ldots, n$, i.i.d.

- Prediction as **linear** functions $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathcal{H} = \mathbb{R}^d$
  
  - Optimal prediction $\theta_\ast \in \mathcal{H}$ minimizing $F(\theta) = \mathbb{E}(y - \langle \theta, \Phi(x) \rangle)^2$
  
  - Assumption: $\|\Phi(x)\| \leq R$ almost surely
  
  - Assumption: $|y| \leq M$ and $|y - \langle \theta_\ast, \Phi(x) \rangle| \leq \sigma$ almost surely
Least-squares regression in finite dimension

• **Data:** \( n \) observations \((x_i, y_i) \in \mathcal{X} \times \mathbb{R}, i = 1, \ldots, n, \text{ i.i.d.}\)

• Prediction as **linear** functions \( \langle \theta, \Phi(x) \rangle \) of features \( \Phi(x) \in \mathcal{H} = \mathbb{R}^d \)
  
  – Optimal prediction \( \theta_\ast \in \mathcal{H} \) minimizing \( F(\theta) = \mathbb{E}(y - \langle \theta, \Phi(x) \rangle)^2 \)
  
  – Assumption: \( \|\Phi(x)\| \leq R \) almost surely
  
  – Assumption: \( |y| \leq M \) and \( |y - \langle \theta_\ast, \Phi(x) \rangle| \leq \sigma \) almost surely

• **Statistical performance of estimators** \( \hat{\theta} \) defined as \( \mathbb{E}F(\hat{\theta}) - F(\theta_\ast) \)

  – Finite dimension: optimal rate \( \frac{\sigma^2 \text{dim}(\mathcal{H})}{n} = \frac{\sigma^2 d}{n} \)
  
  – Attained by empirical risk minimization (ERM) and SGD
Least-squares regression in finite dimension

• **Data:** $n$ observations $(x_i, y_i) \in X \times \mathbb{R}$, $i = 1, \ldots, n$, i.i.d.

• **Prediction as linear functions** $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathcal{H} = \mathbb{R}^d$
  
  – Optimal prediction $\theta^* \in \mathcal{H}$ minimizing $F(\theta) = \mathbb{E}(y - \langle \theta, \Phi(x) \rangle)^2$
  
  – Assumption: $\|\Phi(x)\| \leq R$ almost surely
  
  – Assumption: $|y| \leq M$ and $|y - \langle \theta^*, \Phi(x) \rangle| \leq \sigma$ almost surely

• **Statistical performance of estimators** $\hat{\theta}$ defined as $\mathbb{E}F(\hat{\theta}) - F(\theta^*)$
  
  – Finite dimension: optimal rate $\frac{\sigma^2 \dim(\mathcal{H})}{n} = \frac{\sigma^2 d}{n}$
  
  – Attained by empirical risk minimization (ERM) and SGD

• **What if** $n \gg \dim(\mathcal{H})$?
  
  – Needs assumptions on $\Sigma = \mathbb{E}[\Phi(x) \otimes \Phi(x)]$ and $\theta^*$
**Spectrum of covariance matrix** \( \Sigma = \mathbb{E}[\Phi(x) \otimes \Phi(x)] \)

- **Eigenvalues** \( \lambda_m(\Sigma) \) (in decreasing order)

- **Example**: News dataset \((d = 1,300,000, n = 20,000)\)
Spectrum of covariance matrix \( \Sigma = \mathbb{E} [ \Phi(x) \otimes \Phi(x) ] \)

- **Eigenvalues** \( \lambda_m(\Sigma) \) (in decreasing order)

- **Example**: *News* dataset \((d = 1\ 300\ 000, \ n = 20\ 000)\)

- **Assumption**: \( \text{tr}(\Sigma^{1/\alpha}) = \sum_{m \geq 1} \lambda_m(\Sigma)^{1/\alpha} \) is “small” (compared to \( n \))
  - “Equivalent” to \( \lambda_m(\Sigma) = O(m^{-\alpha}) \)
Difficulty of the learning problem

- Measuring difficulty through “the” norm of $\theta_*$

- Assumption: $\|\sum^{1/2-r} \theta_*\|$ is “small” (compared to $n$)
Difficulty of the learning problem

- Measuring difficulty through “the” norm of $\theta_*$

- **Assumption:** $\|\Sigma^{1/2-r}\theta_*\|$ is “small” (compared to $n$)
  - $r = 1/2$: usual assumption on $\|\theta_*\|$  
  - Larger $r$: simpler problems
  - Smaller $r$: harder problems ($r = 0$ always true)
Difficulty of the learning problem

- Measuring difficulty through “the” norm of $\theta_*$

- **Assumption:** $\|\sum^{1/2} r \theta_*\|$ is “small” (compared to $n$)
  - $r = 1/2$: usual assumption on $\|\theta_*\|$  
  - Larger $r$: simpler problems  
  - Smaller $r$: harder problems ($r = 0$ always true)
- **Easy problems** $r \geq \frac{\alpha - 1}{2\alpha}$: optimal rate is $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$.
Optimal statistical performance

- Easy problems $r \geq \frac{\alpha - 1}{2\alpha}$: optimal rate is $O(n^{-\frac{2r\alpha}{2r\alpha+1}})$, achieved by:
  - Regularized ERM (Caponnetto and De Vito, 2007)
  - Early-stopped gradient descent (Yao et al., 2007)
  - Single-pass averaged SGD (Dieuleveut and Bach, 2016)
Optimal statistical performance

- **Easy problems** \( r \geq \frac{\alpha - 1}{2\alpha} \): optimal rate is \( O(n^{-2r\alpha}) \)

- **Hard problems** \( r \leq \frac{\alpha - 1}{2\alpha} \)
  - Lower bound: \( O(n^{-2r\alpha}) \). Known upper bound: \( O(n^{-2r}) \)
Least-mean-square (LMS) algorithm

- **Least-squares:** \( F(\theta) = \frac{1}{2} \mathbb{E} \left[ (y - \langle \Phi(x), \theta \rangle)^2 \right] \) with \( \theta \in \mathbb{R}^d \)
  - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
  - Iteration: \( \theta_i = \theta_{i-1} - \gamma \left( \langle \Phi(x_i), \theta_{i-1} \rangle - y_i \right) \Phi(x_i) \)
Least-mean-square (LMS) algorithm

- **Least-squares:** \( F(\theta) = \frac{1}{2} \mathbb{E} \left[ (y - \langle \Phi(x), \theta \rangle)^2 \right] \) with \( \theta \in \mathbb{R}^d \)
  
  - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
  
  - Iteration: \( \theta_i = \theta_{i-1} - \gamma \left( \langle \Phi(x_i), \theta_{i-1} \rangle - y_i \right) \Phi(x_i) \)

- **New analysis for averaging and constant step-size** \( \gamma = 1/(4R^2) \)
  
  - Bach and Moulines (2013)
  
  - Assume \( \|\Phi(x)\| \leq R \) and \( |y - \langle \Phi(x), \theta^* \rangle| \leq \sigma \) almost surely
  
  - No assumption regarding lowest eigenvalues of \( \Sigma \)

  - Main result: \( \mathbb{E} F(\bar{\theta}_n) - F(\theta^*) \leq \frac{4\sigma^2 d}{n} + \frac{4R^2 \|\theta_0 - \theta^*\|^2}{n} \)

- **Matches statistical lower bound** (Tsybakov, 2003)
Markov chain interpretation of constant step sizes

- LMS recursion: $\theta_i = \theta_{i-1} - \gamma \langle \Phi(x_i), \theta_{i-1} \rangle - y_i \rangle \Phi(x_i)$
Markov chain interpretation of constant step sizes

- LMS recursion: \( \theta_i = \theta_{i-1} - \gamma \left( \langle \Phi(x_i), \theta_{i-1} \rangle - y_i \right) \Phi(x_i) \)

- The sequence \((\theta_i)_i\) is a homogeneous Markov chain
  - convergence to a stationary distribution \(\pi_\gamma\)
  - with expectation \(\bar{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(d\theta)\)
Markov chain interpretation of constant step sizes

• LMS recursion: $\theta_i = \theta_{i-1} - \gamma (\langle \Phi(x_i), \theta_{i-1} \rangle - y_i) \Phi(x_i)$

• The sequence $(\theta_i)_i$ is a homogeneous Markov chain
  – convergence to a stationary distribution $\pi_\gamma$
  – with expectation $\overline{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma (d\theta)$

• For least-squares, $\overline{\theta}_\gamma = \theta_*$
Markov chain interpretation of constant step sizes

- LMS recursion: \( \theta_i = \theta_{i-1} - \gamma (\langle \Phi(x_i), \theta_{i-1} \rangle - y_i) \Phi(x_i) \)

- The sequence \((\theta_i)_i\) is a \textbf{homogeneous Markov chain}
  - convergence to a stationary distribution \(\pi_\gamma\)
  - with expectation \(\bar{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(d\theta)\)

- \textbf{For least-squares,} \(\bar{\theta}_\gamma = \theta_*\)
Markov chain interpretation of constant step sizes

• LMS recursion: $\theta_i = \theta_{i-1} - \gamma(\langle \Phi(x_i), \theta_{i-1} \rangle - y_i)\Phi(x_i)$

• The sequence $(\theta_i)_i$ is a homogeneous Markov chain
  - convergence to a stationary distribution $\pi_\gamma$
  - with expectation $\bar{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

• For least-squares, $\bar{\theta}_\gamma = \theta_*$
  - $\theta_n$ does not converge to $\theta_*$ but oscillates around it

• Ergodic theorem:
  - Averaged iterates converge to $\bar{\theta}_\gamma = \theta_*$ at rate $O(1/n)$
  - See Dieuleveut, Durmus, and Bach (2017) for more details
Simulations - synthetic examples

- Gaussian distributions - \( d = 20 \)
Simulations - benchmarks

- alpha ($d = 500, n = 500\,000$), news ($d = 1\,300\,000, n = 20\,000$)
Optimal bounds for least-squares?

• Least-squares: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  – What if $d \gg n$?
Optimal bounds for least-squares?

- **Least-squares**: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  - What if $d \gg n$?

- Needs assumptions on $\Sigma = \mathbb{E} [\Phi(x) \otimes \Phi(x)]$ and $\theta^*$
Finer assumptions (Dieuleveut and Bach, 2016)

- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues $\lambda_m$ less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\text{tr } \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small

![Graph showing log(\lambda_m) vs log(m)]
Finer assumptions (Dieuleveut and Bach, 2016)

- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues $\lambda_m$ less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\text{tr } \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\frac{\sigma^2 d}{n}$ by $\frac{\sigma^2 (\gamma n)^{1/\alpha} \text{tr } \Sigma^{1/\alpha}}{n}$
Finer assumptions (Dieuleveut and Bach, 2016)

- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues $\lambda_m$ less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\text{tr } \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\frac{\sigma^2 d}{n}$ by $\frac{\sigma^2 (\gamma n)^{1/\alpha} \text{tr } \Sigma^{1/\alpha}}{n}$

- **Optimal predictor**
  - Pessimistic assumption: $\|\theta_0 - \theta_*\|^2$ finite/small
  - Finer assumption: $\|\Sigma^{1/2-r}(\theta_0 - \theta_*)\|_2$ small, for $r \in [0, 1]$
  - Always satisfied for $r = 0$ and $\theta_0 = 0$, since $\|\Sigma^{1/2} \theta_*\| \leq 2\sqrt{\mathbb{E}y_n^2}$
Finer assumptions (Dieuleveut and Bach, 2016)

- Covariance eigenvalues
  - Pessimistic assumption: all eigenvalues $\lambda_m$ less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\text{tr } \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\frac{\sigma^2 d}{n}$ by $\frac{\sigma^2 (\gamma n)^{1/\alpha} \text{tr } \Sigma^{1/\alpha}}{n}$

- Optimal predictor
  - Pessimistic assumption: $\|\theta_0 - \theta_*\|^2$ finite/small
  - Finer assumption: $\|\Sigma^{1/2-r}(\theta_0 - \theta_*)\|_2$ small, for $r \in [0, 1]$
  - Always satisfied for $r = 0$ and $\theta_0 = 0$, since $\|\Sigma^{1/2}\theta_*\| \leq 2\sqrt{\mathbb{E}y_n^2}$
  - New result: replace $\frac{\|\theta_0 - \theta_*\|^2}{\gamma n}$ by $\frac{\|\Sigma^{1/2-r}(\theta_0 - \theta_*)\|^2}{\gamma^2 r n^{2r}}$. 
Optimal bounds for least-squares?

- **Least-squares**: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  - What if $d \gg n$?

- **Refined assumptions with adaptivity** (Dieuleveut and Bach, 2016)
  - Beyond strong convexity or lack thereof

\[
\mathbb{E} F(\bar{\theta}_n) - F(\theta^*) \leq \inf_{\alpha \geq 1, r \in [0,1]} \frac{4\sigma^2 \text{tr} \Sigma^{1/\alpha}}{n} (\gamma n)^{1/\alpha} + \frac{4\|\Sigma^{1/2 - r} \theta^*\|^2}{\gamma^2 r n^{2r}}
\]

- Previous results: $\alpha = +\infty$ and $r = 1/2$
Optimal bounds for least-squares?

- **Least-squares**: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  - What if $d \gg n$?

- **Refined assumptions with adaptivity** (Dieuleveut and Bach, 2016)
  - Beyond strong convexity or lack thereof

\[
\mathbb{E} F(\bar{\theta}_n) - F(\theta_*) \leq \inf_{\alpha \geq 1, r \in [0,1]} \frac{4\sigma^2 \text{tr} \Sigma^{1/\alpha}}{n} (\gamma n)^{1/\alpha} + \frac{4\|\Sigma^{1/2-r} \theta_*\|_2^2}{\gamma^2 r n^{2r}}
\]

- Previous results: $\alpha = +\infty$ and $r = 1/2$
- Optimal step-size $\gamma$ potentially decaying with $n$, but depends on usually unknown quantities $\alpha$ and $r \Leftrightarrow \text{no adaptivity (yet)}$
Optimal bounds for least-squares?

- **Least-squares**: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  - What if $d \gg n$?

- **Refined assumptions with adaptivity** (Dieuleveut and Bach, 2016)
  - Beyond strong convexity or lack thereof

$$
\mathbb{E} F(\bar{\theta}_n) - F(\theta_*) \leq \inf_{\alpha \geq 1, r \in [0,1]} \frac{4\sigma^2 \text{tr} \Sigma^{1/\alpha}}{n} (\gamma n)^{1/\alpha} + \frac{4\|\Sigma^{1/2-r} \theta_*\|^2}{\gamma^2 r n^{2r}}
$$

  - Previous results: $\alpha = +\infty$ and $r = 1/2$
  - Optimal step-size $\gamma$ potentially decaying with $n$, but depends on usually unknown quantities $\alpha$ and $r \Leftrightarrow$ no adaptivity (yet)
  - Extension to non-parametric estimation (using kernels) with optimal rates when $r \geq (\alpha - 1)/(2\alpha)$, still with $O(n^2)$ running-time
From least-squares to non-parametric estimation

- Extension to Hilbert spaces: \( \Phi(x), \theta \in \mathcal{H} \)

\[
\theta_i = \theta_{i-1} - \gamma \left( \langle \Phi(x_i), \theta_{i-1} \rangle - y_i \right) \Phi(x_i)
\]

- If \( \theta_0 = 0 \), \( \theta_i \) is a linear combination of \( \Phi(x_1), \ldots, \Phi(x_i) \)

\[
\theta_i = \sum_{k=1}^{i} a_k \Phi(x_k) \quad \text{and} \quad a_i = -\gamma \sum_{k=1}^{i-1} a_k \langle \Phi(x_k), \Phi(x_i) \rangle + \gamma y_i
\]
From least-squares to non-parametric estimation

- **Extension to Hilbert spaces:** $\Phi(x), \theta \in \mathcal{H}$

\[
\theta_i = \theta_{i-1} - \gamma \left( \langle \Phi(x_i), \theta_{i-1} \rangle - y_i \right) \Phi(x_i)
\]

- **If $\theta_0 = 0$, $\theta_i$ is a linear combination of $\Phi(x_1), \ldots, \Phi(x_i)$**

\[
\theta_i = \sum_{k=1}^{i} a_k \Phi(x_k) \quad \text{and} \quad a_i = -\gamma \sum_{k=1}^{i-1} a_k \langle \Phi(x_k), \Phi(x_i) \rangle + \gamma y_i
\]

- **Kernel trick:** $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$
  - Reproducing kernel Hilbert spaces and non-parametric estimation
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004); Dieuleveut and Bach (2016)
  - Still $O(n^2)$ overall running-time
Example: Sobolev spaces in one dimension

- $\mathcal{X} = [0, 1]$, functions represented through their Fourier series
  - Weighted Fourier basis $\Phi(x)_m = \lambda_m^{1/2} \cos(2m\pi x)$ (plus sines)
  - Kernel $k(x, x') = \sum_m \lambda_m \cos[2m\pi(x - x')]$
Example: Sobolev spaces in one dimension

- $X = [0, 1]$, functions represented through their Fourier series
  - Weighted Fourier basis $\Phi(x)_m = \lambda_m^{1/2} \cos(2m\pi x)$ (plus sines)
  - Kernel $k(x, x') = \sum_m \lambda_m \cos[2m\pi(x - x')]$

- $\lambda_m \propto m^{-\alpha}$ corresponds to Sobolev penalty on $f_\theta(x) = \langle \theta, \Phi(x) \rangle$

\[
\|f_\theta\|^2 = \|\theta\|^2 = \sum_m |\text{Fourier}(f_\theta)_m|^2 \lambda_m^{-1} \propto \int_0^1 |f_\theta^{(\alpha/2)}(x)|^2 dx
\]
Example: Sobolev spaces in one dimension

- $\mathcal{X} = [0, 1]$, functions represented through their Fourier series
  - Weighted Fourier basis $\Phi(x)_m = \lambda_m^{1/2} \cos(2m\pi x)$ (plus sines)
  - kernel $k(x, x') = \sum_m \lambda_m \cos[2m\pi(x - x')]$
- $\lambda_m \propto m^{-\alpha}$ corresponds to Sobolev penalty on $f_\theta(x) = \langle \theta, \Phi(x) \rangle$
  \[
  \|f_\theta\|^2 = \|\theta\|^2 = \sum_m |\text{Fourier}(f_\theta)_m|^2 \lambda_m^{-1} \propto \int_0^1 |f_\theta^{(\alpha/2)}(x)|^2 dx
  \]
- Adapted norm $\|\sum^{1/2-r} \theta\|^2$ depends on regularity of $f_\theta$
  - $\|\sum^{1/2-r} \theta\|^2 = \sum_m |\text{Fourier}(f_\theta)_m|^2 \lambda_m^{-2r} \propto \int_0^1 |f_\theta^{(r\alpha)}(x)|^2 dx$
  - Optimal rate is $O(n^{-2r\alpha+1})$
New assumption needed

- **Assumption:** \( \| \sum^{\mu/2-1/2} \Phi(x) \| \) almost surely “small”
  - Already used by Steinwart et al. (2009)
  - True for \( \mu = 1 \)
  - Usually \( \mu \geq 1/\alpha \) (equal for Sobolev spaces)
  - Relationship between \( L_\infty \) norm \( \| \cdot \|_{L_\infty} \) and RKHS norm \( \| \cdot \| \)
    \[
    \| g \|_{L_\infty} = O(\| g \|^\mu \| g \|_{L_2}^{1-\mu})
    \]
  - NB: implies bounded leverage scores (Rudi et al., 2015)
Multiple pass SGD (sampling with replacement)

- Algorithm from $n$ i.i.d. observations $(x_i, y_i), i = 1, \ldots, n$:

  $$
  \theta_u = \theta_{u-1} + \gamma(y_i(u) - \langle \theta_{u-1}, \Phi(x_i(u)) \rangle) \Phi(x_i(u))
  $$

- $\bar{\theta}_t$ averaged iterate after $t \geq n$ iterations
Multiple pass SGD (sampling with replacement)

- **Algorithm** from $n$ i.i.d. observations $(x_i, y_i)$, $i = 1, \ldots, n$:
  \[
  \theta_u = \theta_{u-1} + \gamma(y_{i(u)} - \langle \theta_{u-1}, \Phi(x_{i(u)}) \rangle) \Phi(x_{i(u)})
  \]
  - $\bar{\theta}_t$ averaged iterate after $t \geq n$ iterations

- **Theorem** (Pillaud-Vivien, Rudi, and Bach, 2018): Assume $r \leq \frac{\alpha-1}{2\alpha}$.
  - If $\mu \leq 2r$, then after $t = \Theta(n^{\alpha/(2r\alpha+1)})$ iterations, we have:
    \[
    \mathbb{E}F(\bar{\theta}_t) - F(\theta^*) = O(n^{-2r\alpha/(2r\alpha+1)})
    \]
  - Otherwise, then after $t = \Theta(n^{1/\mu} (\log n)^{\frac{1}{\mu}})$ iterations, we have:
    \[
    \mathbb{E}F(\bar{\theta}_t) - F(\theta^*) \leq O(n^{-2r/\mu})
    \]

- Proof technique following Rosasco and Villa (2015)
Proof sketch

- **Algorithm** from \( n \) i.i.d. observations \((x_i, y_i), i = 1, \ldots, n:\)
  \[
  \theta_u = \theta_{u-1} + \gamma (y_i(u) - \langle \theta_{u-1}, \Phi(x_i(u)) \rangle ) \Phi(x_i(u))
  \]
  \( \bar{\theta}_t \) averaged iterate after \( t \geq n \) iterations

- Following Rosasco and Villa (2015), consider batch gradient recursion
  \[
  \eta_u = \eta_{u-1} + \frac{\gamma}{n} \sum_{i=1}^{n} (y_i - \langle \eta_{u-1}, \Phi(x_i) \rangle ) \Phi(x_i)
  \]
  \( \bar{\eta}_t \) averaged iterate after \( t \geq n \) iterations
Proof sketch

• **Algorithm** from \( n \) i.i.d. observations \((x_i, y_i), i = 1, \ldots, n:\)

\[
\theta_u = \theta_{u-1} + \gamma (y_{i(u)} - \langle \theta_{u-1}, \Phi(x_{i(u)}) \rangle) \Phi(x_{i(u)})
\]

\(- \bar{\theta}_t \) averaged iterate after \( t \geq n \) iterations

• Following Rosasco and Villa (2015), consider batch gradient recursion

\[
\eta_u = \eta_{u-1} + \frac{\gamma}{n} \sum_{i=1}^{n} (y_i - \langle \eta_{u-1}, \Phi(x_i) \rangle) \Phi(x_i)
\]

\(- \bar{\eta}_t \) averaged iterate after \( t \geq n \) iterations

• As long as \( t = O(n^{1/\mu}) \)

  – **Property 1**: \( \mathbb{E} F(\bar{\theta}_t) - \mathbb{E} F(\bar{\eta}_t) = O\left(\frac{t^{1/\alpha}}{n}\right) \)

  – **Property 2**: \( \mathbb{E} F(\bar{\eta}_t) - F(\theta^*) = O\left(\frac{t^{1/\alpha}}{n}\right) + O(t^{-2r}) \)
Multiple pass SGD (sampling with replacement)

- **Algorithm** from $n$ i.i.d. observations $(x_i, y_i), i = 1, \ldots, n$:

  $$\theta_u = \theta_{u-1} + \gamma \left( y_{i(u)} - \langle \theta_{u-1}, \Phi(x_{i(u)}) \rangle \right) \Phi(x_{i(u)})$$

  $\bar{\theta}_t$ averaged iterate after $t \geq n$ iterations

- **Theorem** (Pillaud-Vivien, Rudi, and Bach, 2018): Assume $r \leq \frac{\alpha - 1}{2\alpha}$.

  - If $\mu \leq 2r$, then after $t = \Theta(n^{\alpha/(2r\alpha+1)})$ iterations, we have:
    $$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) = O\left(n^{-2r\alpha/(2r\alpha+1)}\right)$$
    Optimal

  - Otherwise, then after $t = \Theta(n^{1/\mu} (\log n)^{1/\mu})$ iterations, we have:
    $$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leq O\left(n^{-2r/\mu}\right)$$
    Improved

- **Proof technique** following Rosasco and Villa (2015)
Statistical optimality

- If $\mu \leq 2r$, then after $t = \Theta(n^{\alpha/(2r\alpha+1)})$ iterations, we have:

  $$\mathbb{E} F(\bar{\theta}_t) - F(\theta^*) = O(n^{-2r\alpha/(2r\alpha+1)})$$  \hspace{1cm} \text{Optimal}

- Otherwise, then after $t = \Theta(n^{1/\mu} (\log n)^{1/\mu})$ iterations, we have:

  $$\mathbb{E} F(\bar{\theta}_t) - F(\theta^*) \leq O(n^{-2r/\mu})$$  \hspace{1cm} \text{Improved}
Simulations

- Synthetic examples
  - One-dimensional kernel regression
  - Sobolev spaces
  - Arbitrary chosen values for $r$ and $\alpha$

- Check optimal number of iterations over the data
Simulations

- **Synthetic examples**
  - One-dimensional kernel regression
  - Sobolev spaces
  - Arbitrary chosen values for $r$ and $\alpha$

- **Check optimal number of iterations over the data**

- **Comparing three sampling schemes**
  - With replacement
  - Without replacement (cycling with random reshuffling)
  - Cycling
Simulations (sampling with replacement)

\[ \alpha = \frac{3}{2}, \quad r = \frac{1}{3} > \frac{\alpha - 1}{2\alpha} \]

\[ \alpha = 4, \quad r = \frac{1}{4} = \frac{\alpha - 1}{2\alpha} \]

\[ \alpha = \frac{5}{2}, \quad r = \frac{1}{5} < \frac{\alpha - 1}{2\alpha} \]

\[ \alpha = 3, \quad r = \frac{1}{6} < \frac{\alpha - 1}{2\alpha} \]
Simulations (sampling without replacement)

\[ \alpha = \frac{3}{2}, \quad r = \frac{1}{3} > \frac{\alpha - 1}{2\alpha} \]

\[ \alpha = 4, \quad r = \frac{1}{4} = \frac{\alpha - 1}{2\alpha} \]

\[ \alpha = \frac{5}{2}, \quad r = \frac{1}{5} < \frac{\alpha - 1}{2\alpha} \]

\[ \alpha = 3, \quad r = \frac{1}{6} < \frac{\alpha - 1}{2\alpha} \]
Simulations (cycling)

\[ \alpha = \frac{3}{2}, \, r = \frac{1}{3} > \frac{(\alpha - 1)}{2\alpha} \]

\[ \alpha = 4, \, r = \frac{1}{4} = \frac{(\alpha - 1)}{2\alpha} \]

\[ \alpha = \frac{5}{2}, \, r = \frac{1}{5} < \frac{(\alpha - 1)}{2\alpha} \]

\[ \alpha = 3, \, r = \frac{1}{6} < \frac{(\alpha - 1)}{2\alpha} \]
Simulations - Benchmarks

- MNIST dataset with linear kernel
Conclusion

• Benefits of multiple passes
  – Number of passes grows with sample size for “hard” problems
  – First provable improvement of multiple passes over SGD

[NB: Hardt et al. (2016); Lin and Rosasco (2017) consider small step-sizes]
Conclusion

• **Benefits of multiple passes**
  
  – Number of passes grows with sample size for “hard” problems
  – First provable improvement of multiple passes over SGD
    
    [NB: Hardt et al. (2016); Lin and Rosasco (2017) consider small step-sizes]

• **Current work - Extensions**
  
  – Study of cycling and sampling without replacement
    
    (Shamir, 2016; Gürbüzbalaban et al., 2015)
  – Mini-batches
  – Beyond least-squares
  – Optimal efficient algorithms for the situation $\mu > 2r$
  – Combining analysis with exponential convergence of testing errors (Pillaud-Vivien, Rudi, and Bach, 2017)


