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

Newton Institute, Cambridge - June 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 $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, **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 $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, **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 $F(\theta) = \mathbb{E}(y - \langle \theta, \Phi(x) \rangle)^2$

Least-squares regression in finite dimension

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, **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: $\|\Phi(x)\| \leqslant R$ almost surely
  - Assumption: $|y| \leqslant M$ and $|y - \langle \theta_*, \Phi(x) \rangle| \leqslant \sigma$ almost surely

Least-squares regression in finite dimension

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, **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: $\|\Phi(x)\| \leqslant R$ almost surely
  - Assumption: $|y| \leqslant M$ and $|y - \langle \theta_*, \Phi(x) \rangle| \leqslant \sigma$ 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

Least-squares regression in finite dimension

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

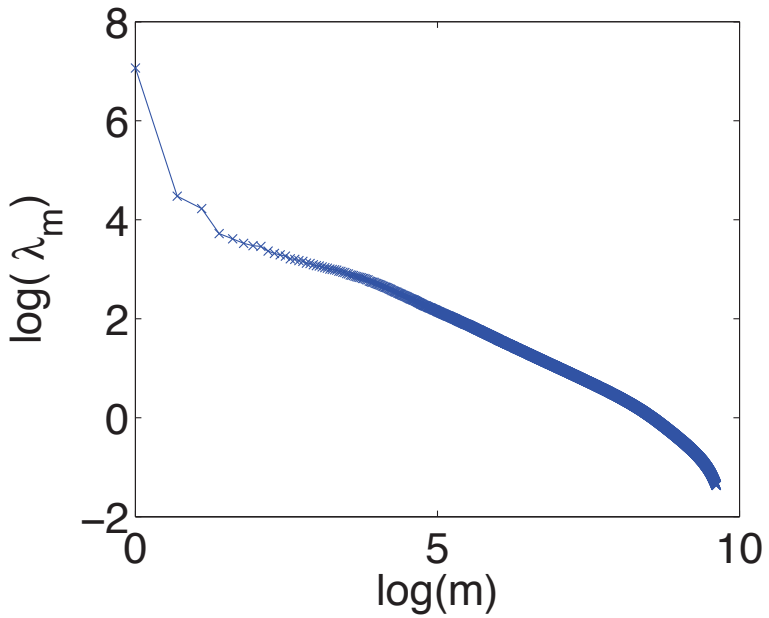
- Prediction as **linear** functions $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathcal{H} = \mathbb{R}^d$
  - Optimal prediction $\theta_* \in \mathcal{H}$ minimizing $F(\theta) = \mathbb{E}(y - \langle \theta, \Phi(x) \rangle)^2$
  - Assumption: $\|\Phi(x)\| \leqslant R$ almost surely
  - Assumption: $|y| \leqslant M$ and $|y - \langle \theta_*, \Phi(x) \rangle| \leqslant \sigma$ 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 $\Sigma = \mathbb{E}\big[\Phi(x) \otimes \Phi(x)\big]$ and θ_*

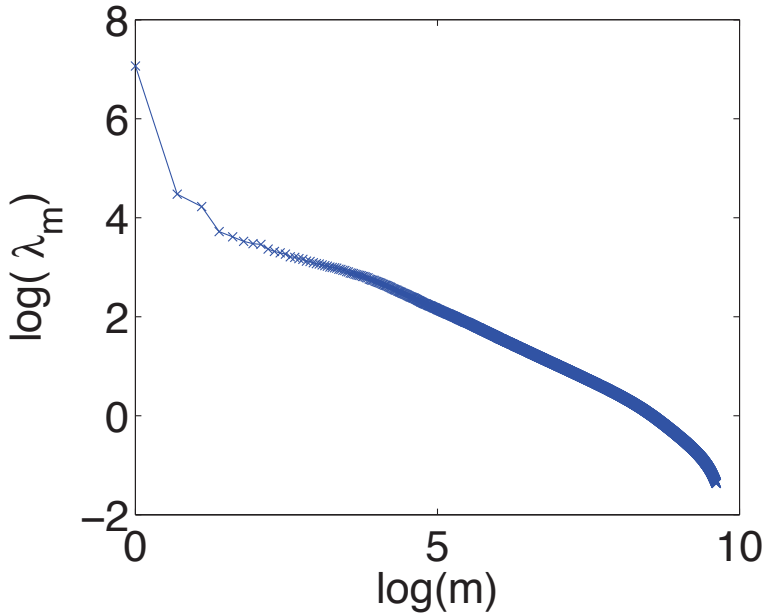
Spectrum of covariance matrix $\Sigma = \mathbb{E}\big[\Phi(x) \otimes \Phi(x)\big]$

- **Eigenvalues** $\lambda_m(\Sigma)$ (in decreasing order)
- **Example**: *News* dataset ($d = 1\ 300\ 000$, $n = 20\ 000$)



Spectrum of covariance matrix $\Sigma = \mathbb{E}\big[\Phi(x) \otimes \Phi(x)\big]$

- **Eigenvalues** $\lambda_m(\Sigma)$ (in decreasing order)
- **Example**: *News* dataset ($d = 1\ 300\ 000$, $n = 20\ 000$)



- **Assumption**: $\mathrm{tr}(\Sigma^{1/\alpha}) = \sum_{m \geqslant 1} \lambda_m(\Sigma)^{1/\alpha}$ is "small" (compared to n)
  - "Equivalent" to $\lambda_m(\Sigma) = O(m^{-\alpha})$

Difficulty of the learning problem

- **Measuring difficulty through "the" norm of θ_***
- **Assumption**: $\|\Sigma^{1/2-r}\theta_*\|$ is "small" (compared to n)

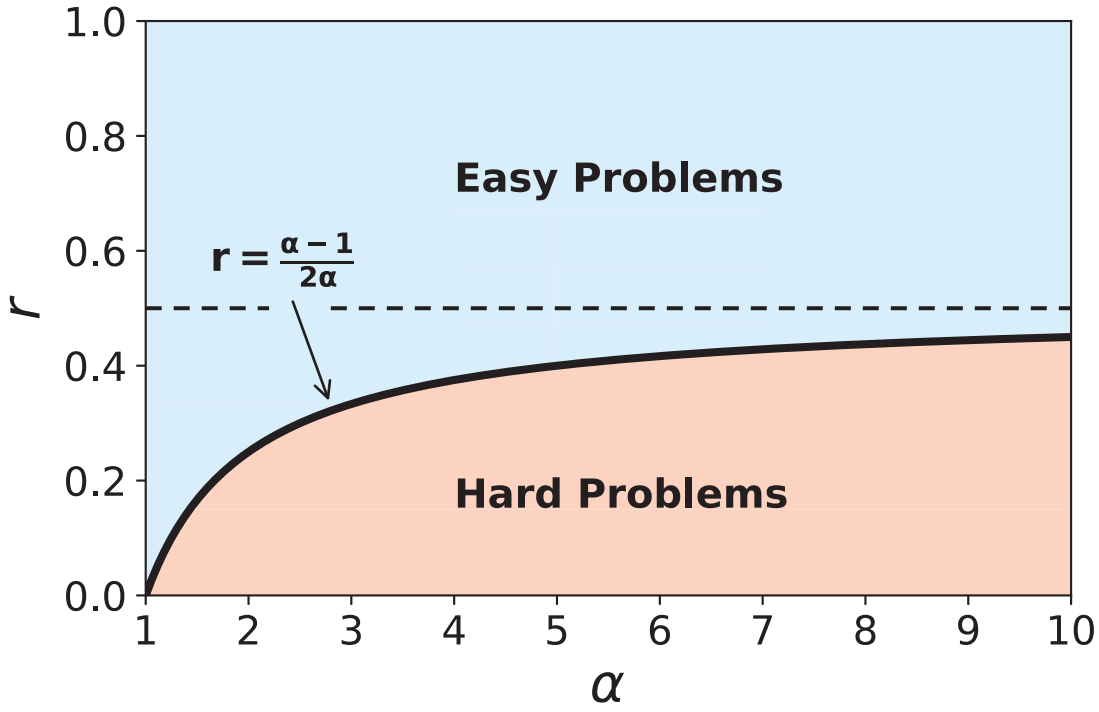
Difficulty of the learning problem

- **Measuring difficulty through "the" norm of θ_***

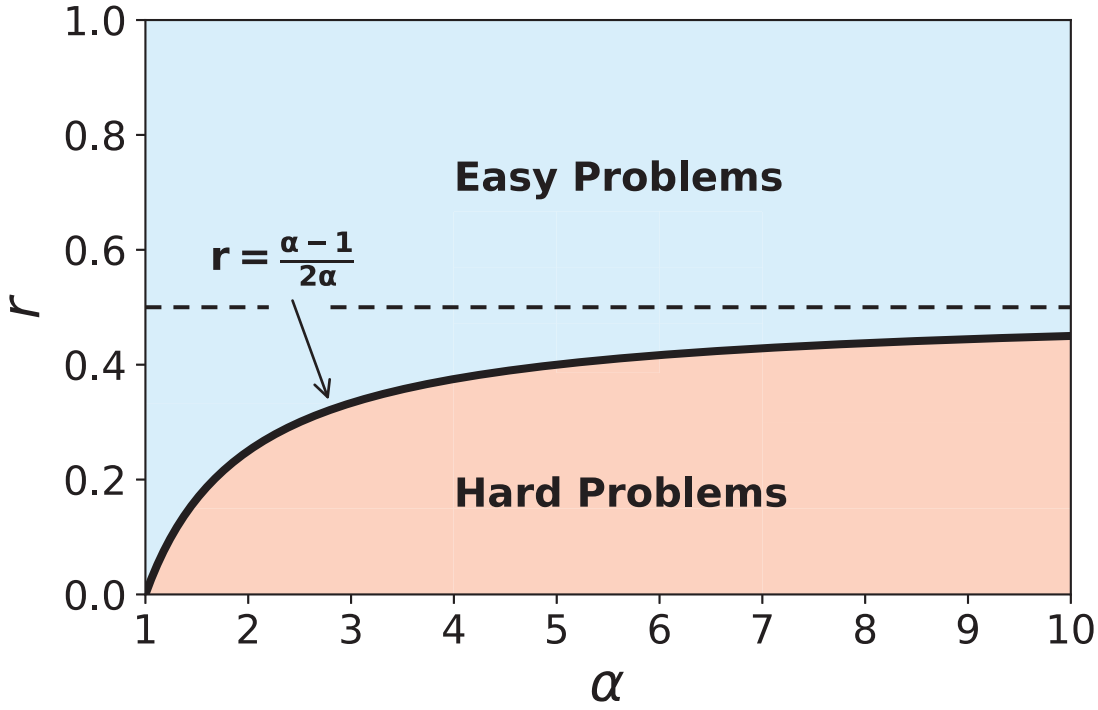
- **Assumption**: $\|\Sigma^{1/2-r}\theta_*\|$ is "small" (compared to n)
  - $r = 1/2$: usual assumption on $\|\theta_*\|$
  - Larger r: simpler problems
  - Smaller r: harder problems ($r = 0$ always true)

Difficulty of the learning problem

- **Measuring difficulty through "the" norm of θ_***
- **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)

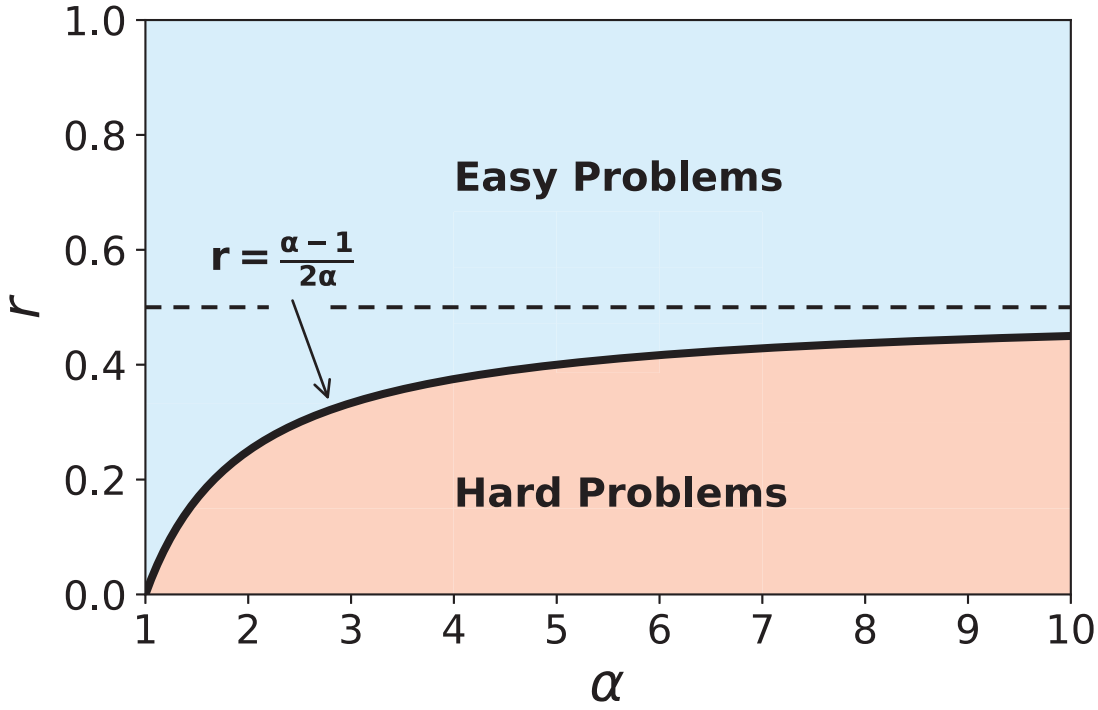


Optimal statistical performance



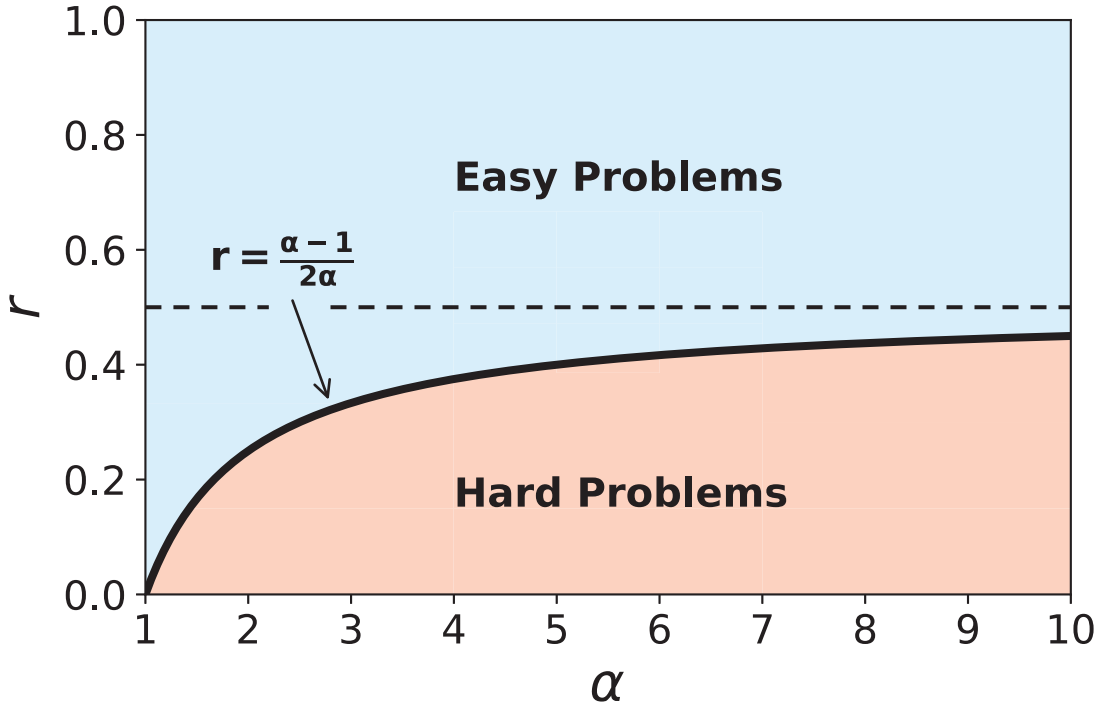
- **Easy problems** $r \geqslant \frac{\alpha-1}{2\alpha}$: optimal rate is $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$

Optimal statistical performance



- **Easy problems** $r \geqslant \frac{\alpha-1}{2\alpha}$: 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** $r \geqslant \frac{\alpha-1}{2\alpha}$: optimal rate is $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$
- **Hard problems** $r \leqslant \frac{\alpha-1}{2\alpha}$
  - Lower bound: $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$. Known upper bound: $O(n^{-2r})$

Least-mean-square (LMS) algorithm

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

Least-mean-square (LMS) algorithm

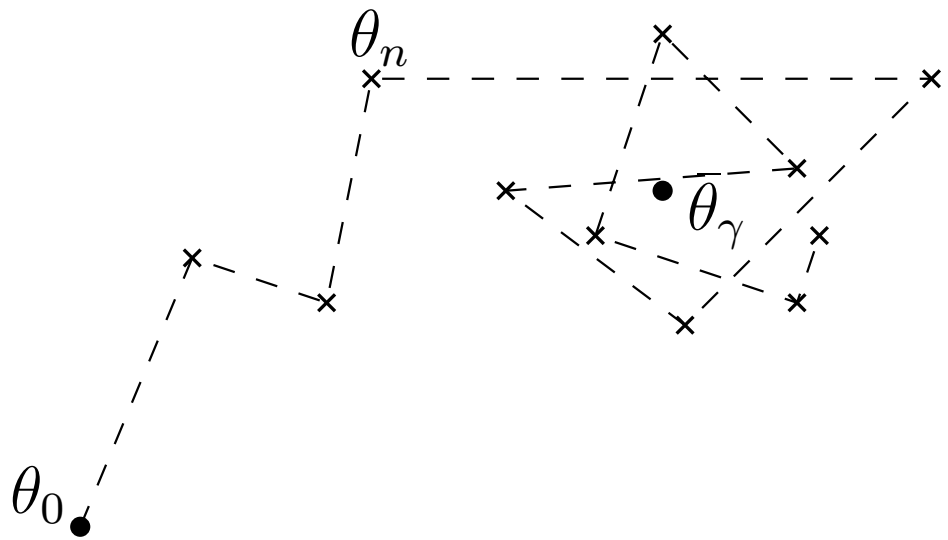
- **Least-squares**: $F(\theta) = \frac{1}{2}\mathbb{E}\big[(y - \langle \Phi(x), \theta \rangle)^2\big]$ with $\theta \in \mathbb{R}^d$
  - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
  - Iteration: $\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1} \rangle - y_i\big)\Phi(x_i)$
- **New analysis for averaging and constant step-size** $\gamma = 1/(4R^2)$
  - Bach and Moulines (2013)
  - Assume $\|\Phi(x)\| \leqslant R$ and $|y - \langle \Phi(x), \theta_* \rangle| \leqslant \sigma$ almost surely
  - No assumption regarding lowest eigenvalues of Σ
  - Main result: $\boxed{\mathbb{E}F(\bar{\theta}_n) - F(\theta_*) \leqslant \dfrac{4\sigma^2 d}{n} + \dfrac{4R^2\|\theta_0 - \theta_*\|^2}{n}}$
- **Matches statistical lower bound** (Tsybakov, 2003)

Markov chain interpretation of constant step sizes

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

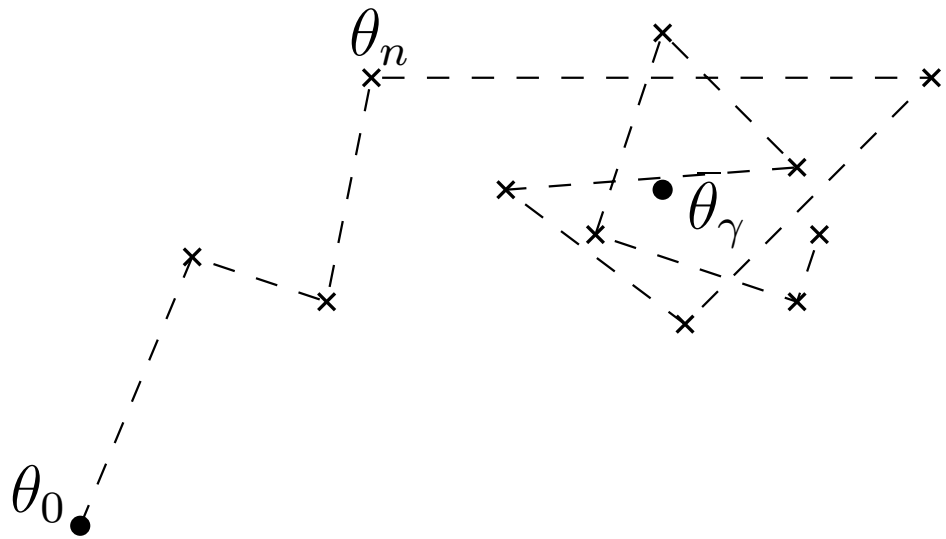
Markov chain interpretation of constant step sizes

- LMS recursion: $\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1}\rangle - y_i\big)\Phi(x_i)$
- The sequence $(\theta_i)_i$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$



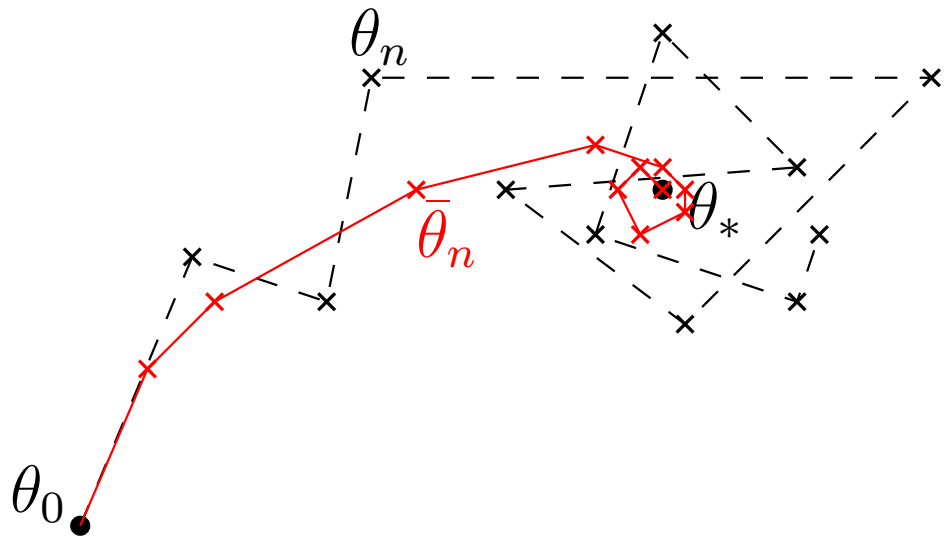
Markov chain interpretation of constant step sizes

- LMS recursion: $\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1}\rangle - y_i\big)\Phi(x_i)$
- The sequence $(\theta_i)_i$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$
- **For least-squares,** $\bar{\theta}_\gamma = \theta_*$



Markov chain interpretation of constant step sizes

- LMS recursion: $\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1}\rangle - y_i\big)\Phi(x_i)$
- The sequence $(\theta_i)_i$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$
- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**

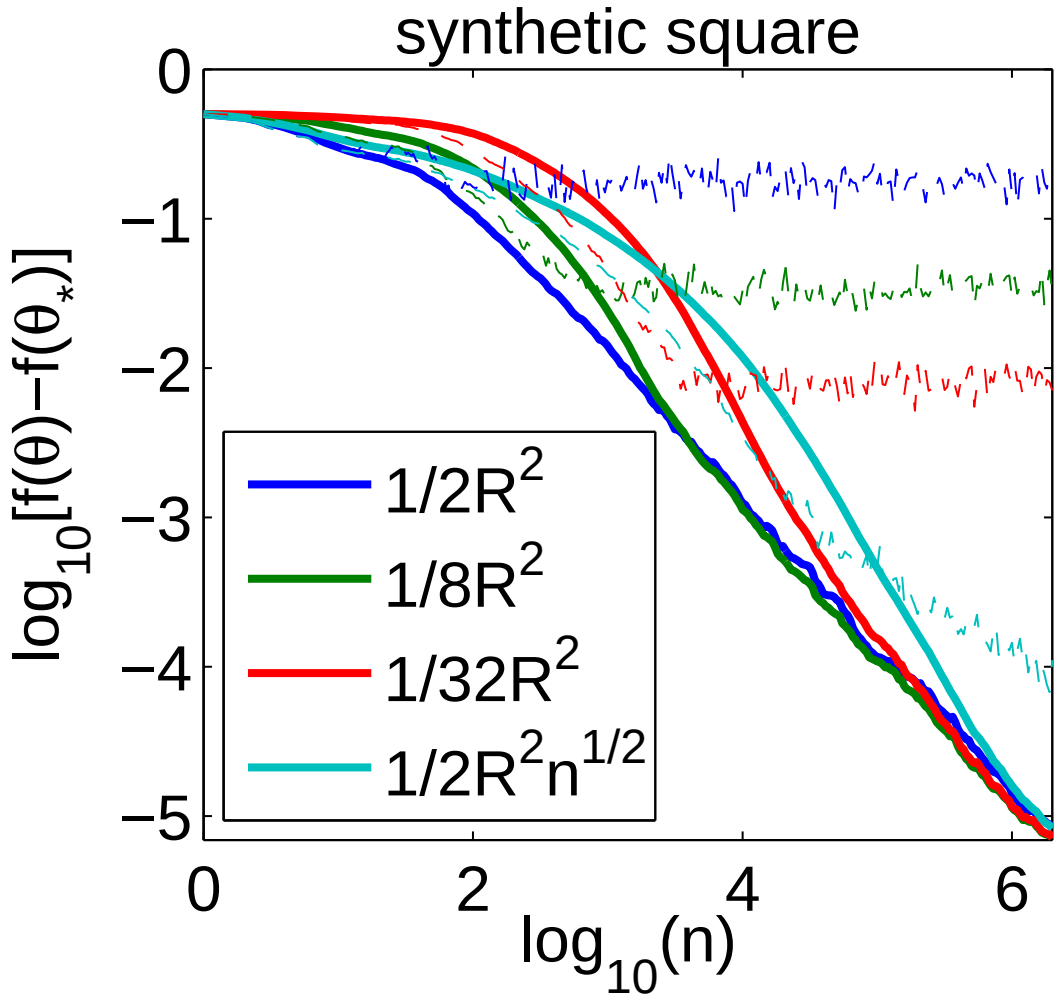


Markov chain interpretation of constant step sizes

- LMS recursion: $\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1}\rangle - y_i\big)\Phi(x_i)$
- The sequence $(\theta_i)_i$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$
- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**
  - θ_n does not converge to θ_* 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

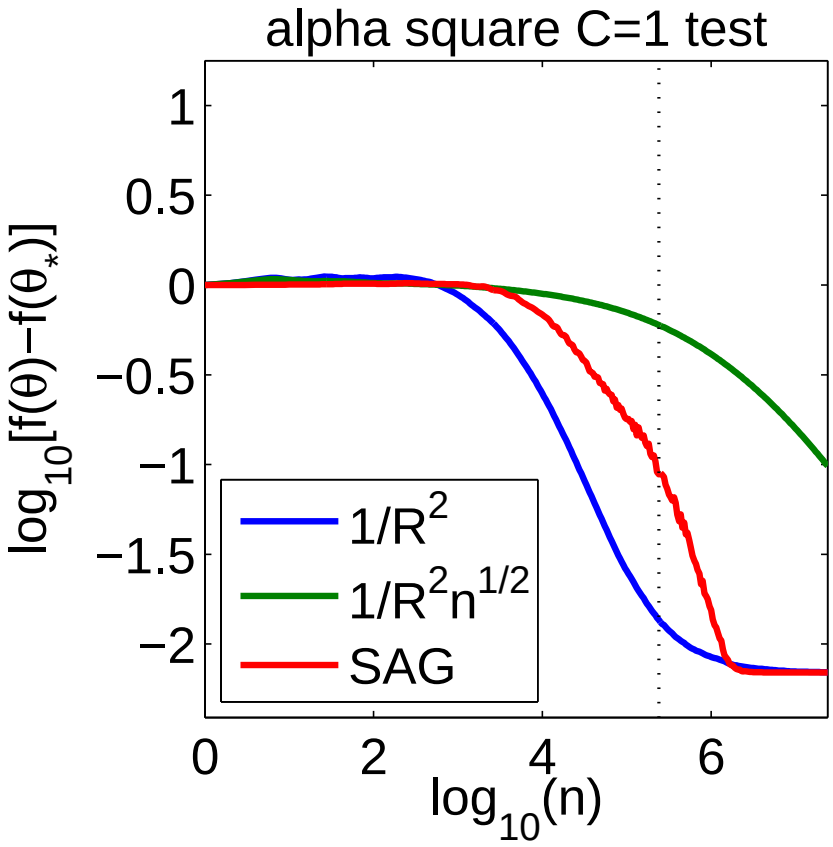
Simulations - synthetic examples

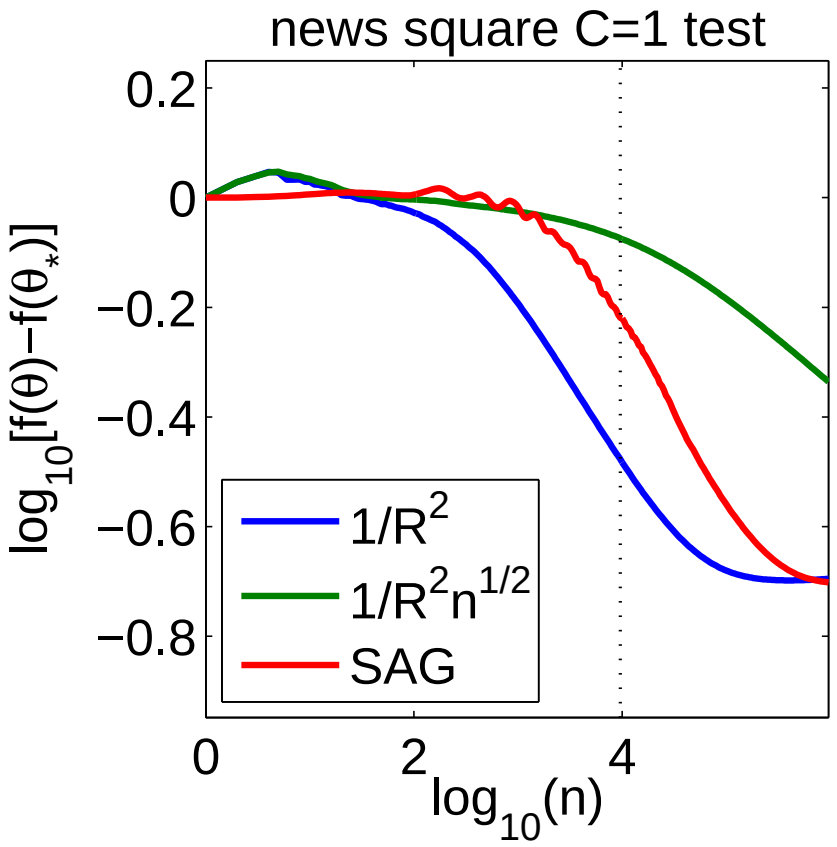
- Gaussian distributions - $d = 20$



Simulations - benchmarks

- *alpha* $(d = 500,\ n = 500\ 000)$, *news* $(d = 1\ 300\ 000,\ n = 20\ 000)$





Optimal bounds for least-squares?

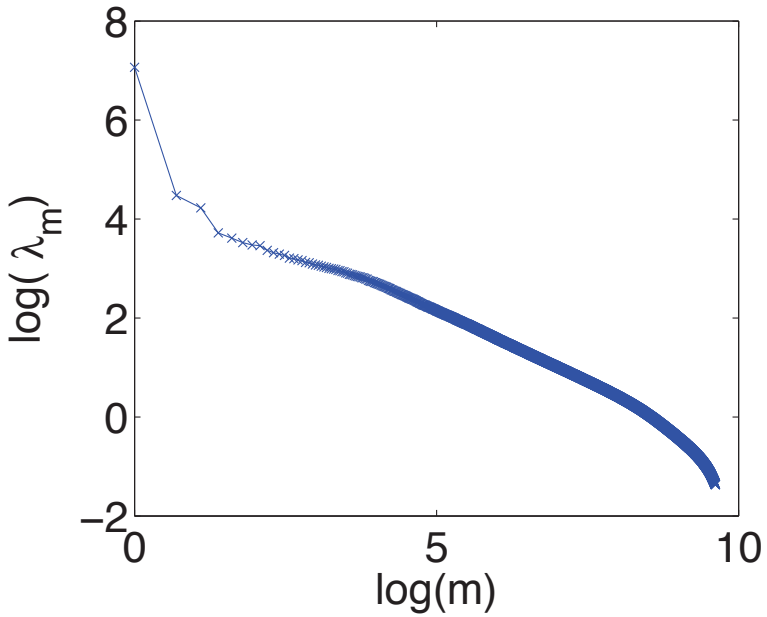
- **Least-squares**: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  - What if $d \gg n$?

Optimal bounds for least-squares?

- **Least-squares**: cannot beat $\sigma^2 d/n$ (Tsybakov, 2003). Really?
  - What if $d \gg n$?
- **Needs assumptions on** $\Sigma = \mathbb{E}\big[\Phi(x) \otimes \Phi(x)\big]$ **and** θ_*

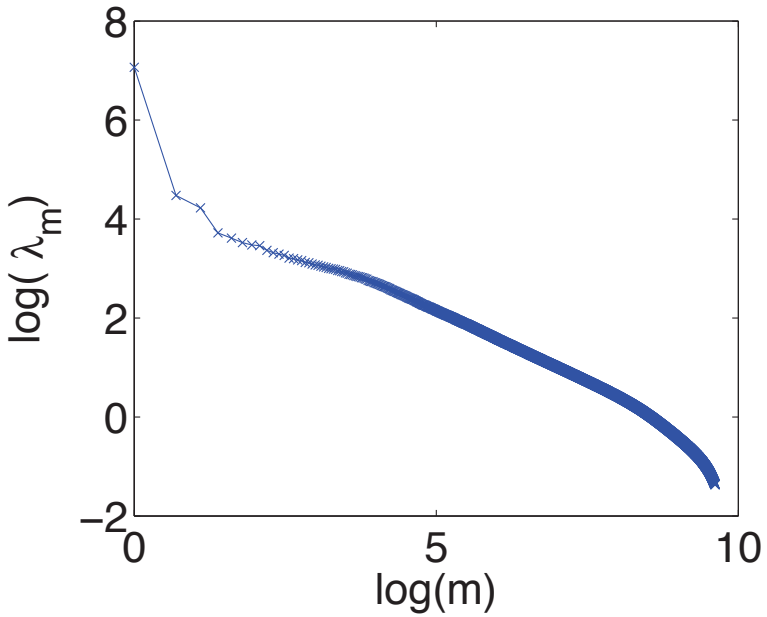
Finer assumptions (Dieuleveut and Bach, 2016)

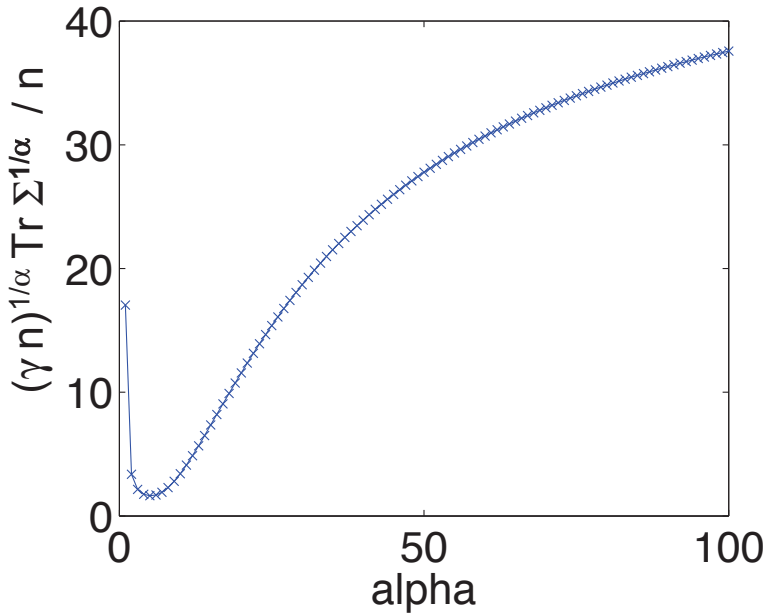
- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues λ_m less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\operatorname{tr} \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small



Finer assumptions (Dieuleveut and Bach, 2016)

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





Finer assumptions (Dieuleveut and Bach, 2016)

- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues λ_m less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\operatorname{tr} \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\dfrac{\sigma^2 d}{n}$ by $\dfrac{\sigma^2 (\gamma n)^{1/\alpha} \operatorname{tr} \Sigma^{1/\alpha}}{n}$
- **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 $\|\Sigma^{1/2}\theta_*\| \leqslant 2\sqrt{\mathbb{E} y_n^2}$

Finer assumptions (Dieuleveut and Bach, 2016)

- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues λ_m less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\operatorname{tr} \Sigma^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\dfrac{\sigma^2 d}{n}$ by $\dfrac{\sigma^2 (\gamma n)^{1/\alpha} \operatorname{tr} \Sigma^{1/\alpha}}{n}$
- **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 $\|\Sigma^{1/2}\theta_*\| \leqslant 2\sqrt{\mathbb{E} y_n^2}$
  - New result: replace $\dfrac{\|\theta_0 - \theta_*\|^2}{\gamma n}$ by $\dfrac{\|\Sigma^{1/2-r}(\theta_0 - \theta_*)\|^2}{\gamma^{2r} n^{2r}}$

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

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

  - 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

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

  - Previous results: $\alpha = +\infty$ and $r = 1/2$
  - Optimal step-size γ potentially decaying with n, but depends on usually unknown quantities α and $r \Leftrightarrow$ 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

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

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

From least-squares to non-parametric estimation

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

$$\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1} \rangle - y_i\big)\Phi(x_i)$$

- **If $\theta_0 = 0$, θ_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**: $\Phi(x), \theta \in \mathcal{H}$

$$\theta_i = \theta_{i-1} - \gamma\big(\langle \Phi(x_i), \theta_{i-1} \rangle - y_i\big)\Phi(x_i)$$

- **If $\theta_0 = 0$, θ_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 $k(x, x') = \sum_m \lambda_m \cos\left[2m\pi(x - x')\right]$

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 $k(x, x') = \sum_m \lambda_m \cos\left[2m\pi(x - x')\right]$

- $\lambda_m \propto m^{-\alpha}$ 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 $k(x, x') = \sum_m \lambda_m \cos\left[2m\pi(x - x')\right]$

- $\lambda_m \propto m^{-\alpha}$ 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 $\|\Sigma^{1/2-r}\theta\|^2$ depends on regularity of f_θ
  - $\|\Sigma^{1/2-r}\theta\|^2 = \sum_m |\text{Fourier}(f_\theta)_m|^2 \lambda_m^{-2r} \propto \int_0^1 |f_\theta^{(r\alpha)}(x)|^2 dx$
  - Optimal rate is $O(n^{\frac{-2r\alpha}{2r\alpha+1}})$

New assumption needed

- **Assumption**: $\|\Sigma^{\mu/2-1/2}\Phi(x)\|$ almost surely "small"
  - Already used by Steinwart et al. (2009)
  - True for $\mu = 1$
  - Usually $\mu \geqslant 1/\alpha$ (equal for Sobolev spaces)
  - Relationship between L_∞ 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), $i = 1, \ldots, n$:

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

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

Multiple pass SGD (sampling with replacement)

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

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

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

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

$$\mathbb{E} F(\bar{\theta}_t) - F(\theta_*) \leqslant O(n^{-2r/\mu})$$

- Proof technique following Rosasco and Villa (2015)

Proof sketch

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

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

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

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

$$\eta_u = \theta_{u-1} + \frac{\gamma}{n}\sum_{i=1}^{n}\big(y_i - \langle \theta_{u-1}, \Phi(x_i)\rangle\big)\Phi(x_i)$$

  - $\bar{\eta}_t$ averaged iterate after $t \geqslant n$ iterations

Proof sketch

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

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

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

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

$$\eta_u = \theta_{u-1} + \frac{\gamma}{n}\sum_{i=1}^{n}\big(y_i - \langle \theta_{u-1}, \Phi(x_i)\rangle\big)\Phi(x_i)$$

  - $\bar{\eta}_t$ averaged iterate after $t \geqslant n$ iterations

- As long as $t = O(n^{1/\mu})$
  - **Property 1**: $\mathbb{E}F(\bar{\theta}_t) - \mathbb{E}F(\bar{\eta}_t) = O\Big(\dfrac{t^{1/\alpha}}{t}\Big)$
  - **Property 2**: $\mathbb{E}F(\bar{\eta}_t) - F(\theta_*) = O\Big(\dfrac{t^{1/\alpha}}{n}\Big) + O(t^{-2r})$

Multiple pass SGD (sampling with replacement)

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

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

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

- **Theorem** (Pillaud-Vivien, Rudi, and Bach, 2018): Assume $r \leqslant \frac{\alpha-1}{2\alpha}$.
  - If $\mu \leqslant 2r$, then after $t = \Theta(n^{\alpha/(2r\alpha+1)})$ iterations, we have:

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

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

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leqslant O(n^{-2r/\mu}) \qquad \text{Improved}$$

- Proof technique following Rosasco and Villa (2015)

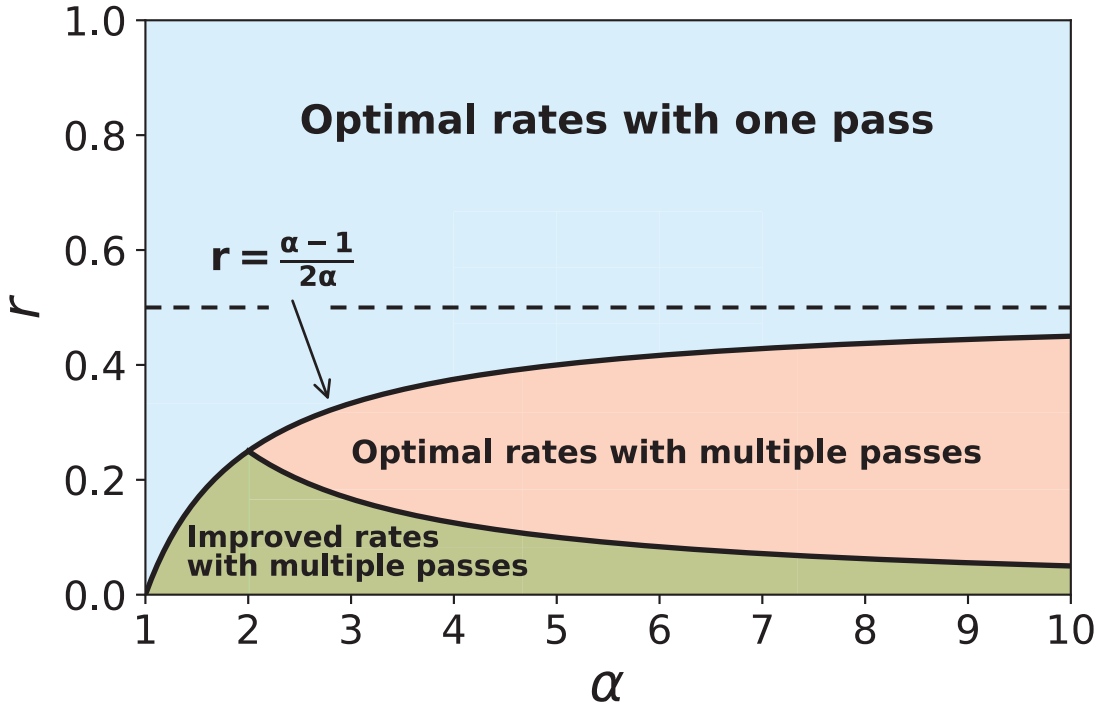
Statistical optimality

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

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

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

$$\mathbb{E}F(\bar{\theta}_t) - F(\theta_*) \leqslant O(n^{-2r/\mu}) \qquad \text{Improved}$$



Simulations

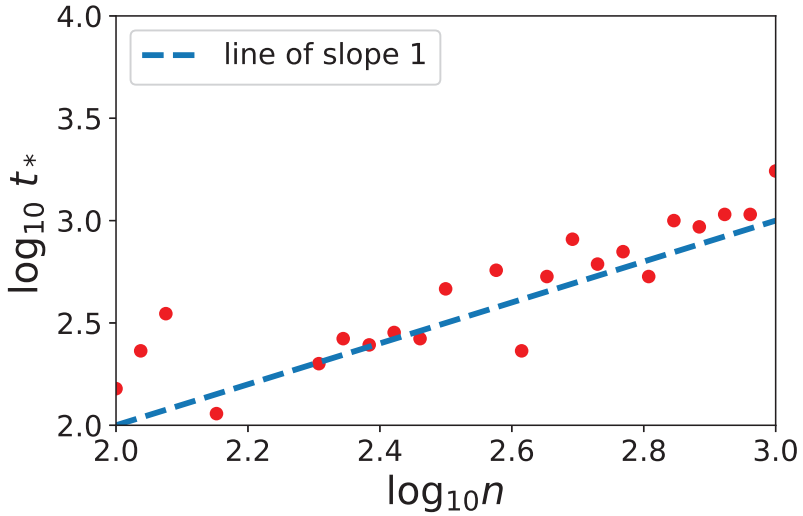
- **Synthetic examples**
  - One-dimensional kernel regression
  - Sobolev spaces
  - Arbitrary chosen values for r and α
- **Check optimal number of iterations over the data**

Simulations

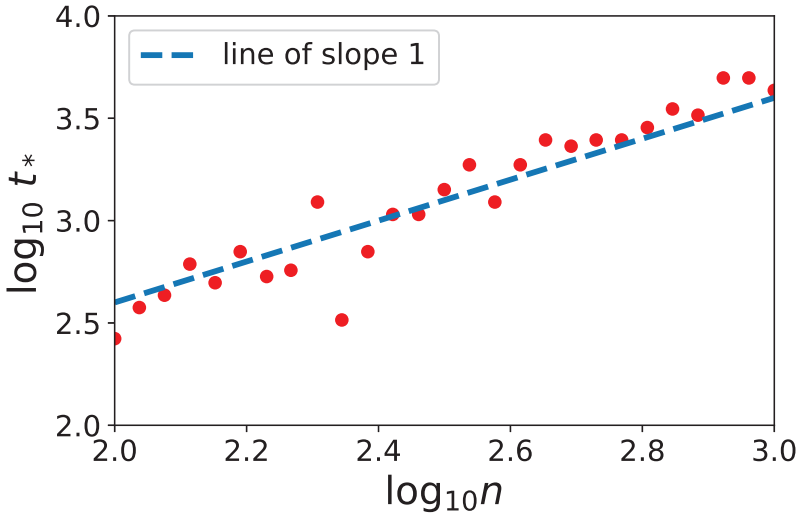
- **Synthetic examples**
  - One-dimensional kernel regression
  - Sobolev spaces
  - Arbitrary chosen values for r and α
- **Check optimal number of iterations over the data**
- **Comparing three sampling schemes**
  - With replacement
  - Without replacement (cycling with random reshuffling)
  - Cycling

Simulations (sampling with replacement)

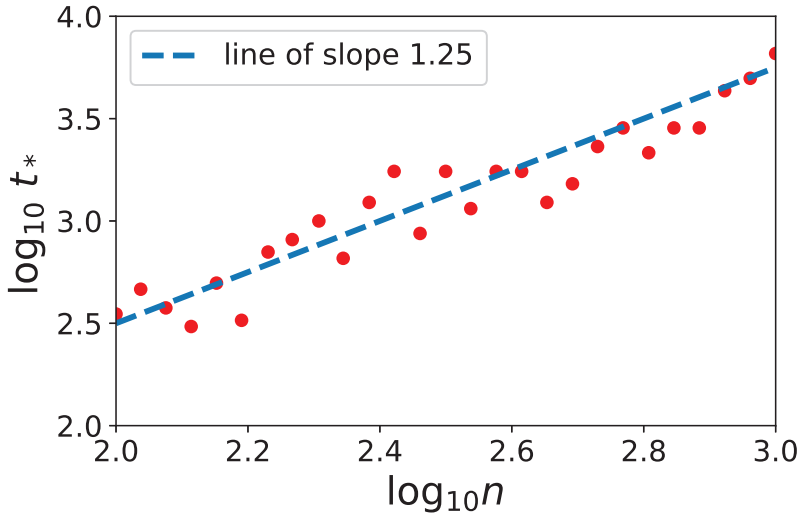
$\alpha = 3/2, r = 1/3 > (\alpha - 1)/(2\alpha)$



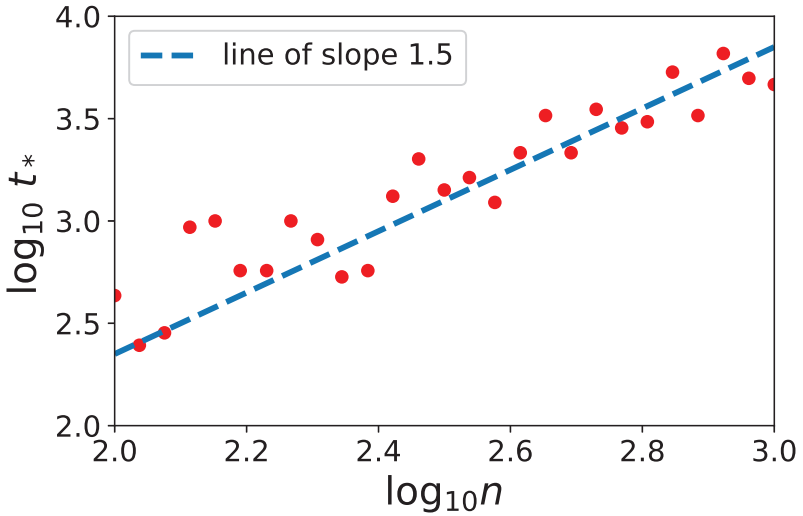
$\alpha = 4, r = 1/4 = (\alpha - 1)/(2\alpha)$



$\alpha = 5/2, r = 1/5 < (\alpha - 1)/(2\alpha)$

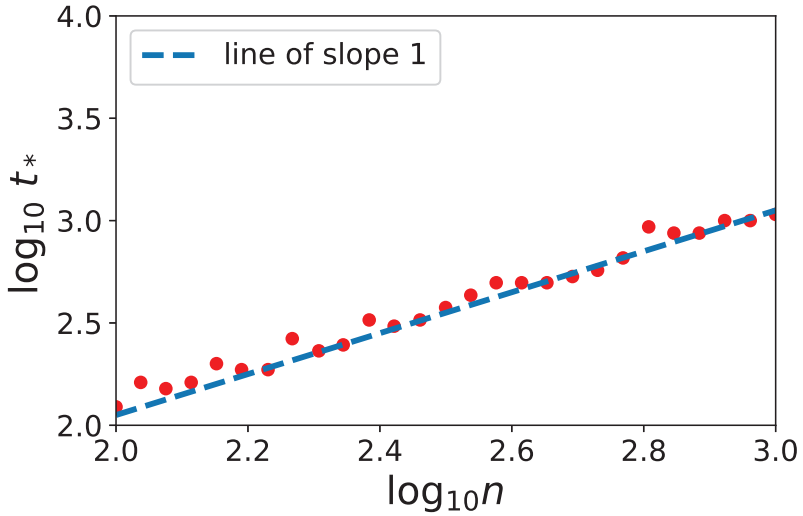


$\alpha = 3, r = 1/6 < (\alpha - 1)/(2\alpha)$

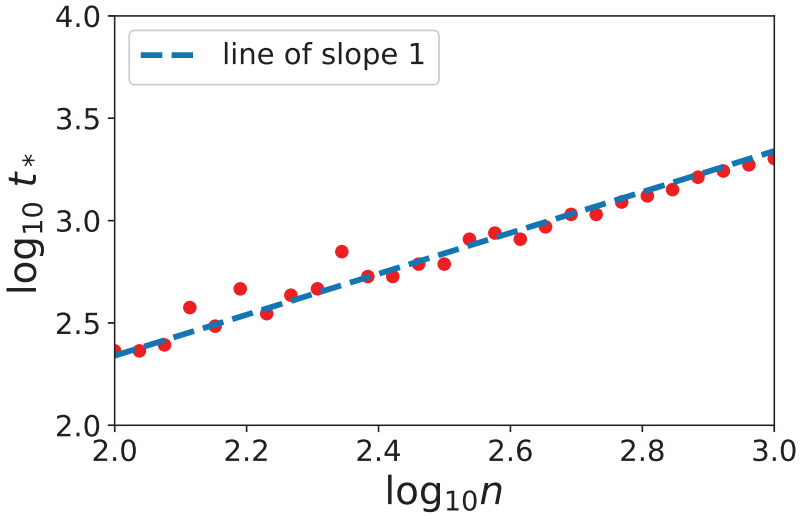


Simulations (sampling without replacement)

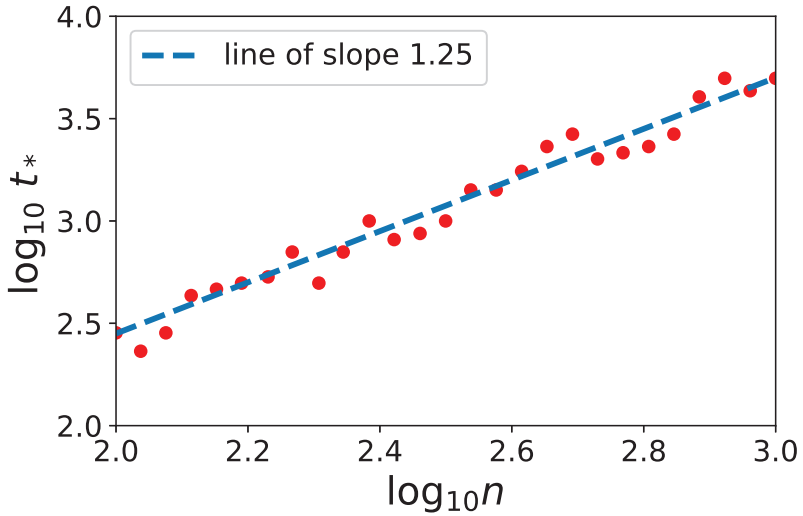
$\alpha = 3/2, r = 1/3 > (\alpha - 1)/(2\alpha)$



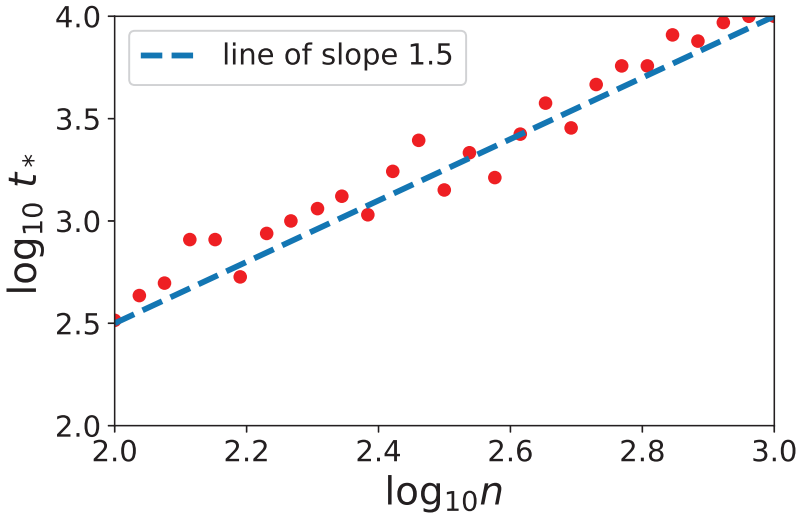
$\alpha = 4, r = 1/4 = (\alpha - 1)/(2\alpha)$



$\alpha = 5/2, r = 1/5 < (\alpha - 1)/(2\alpha)$

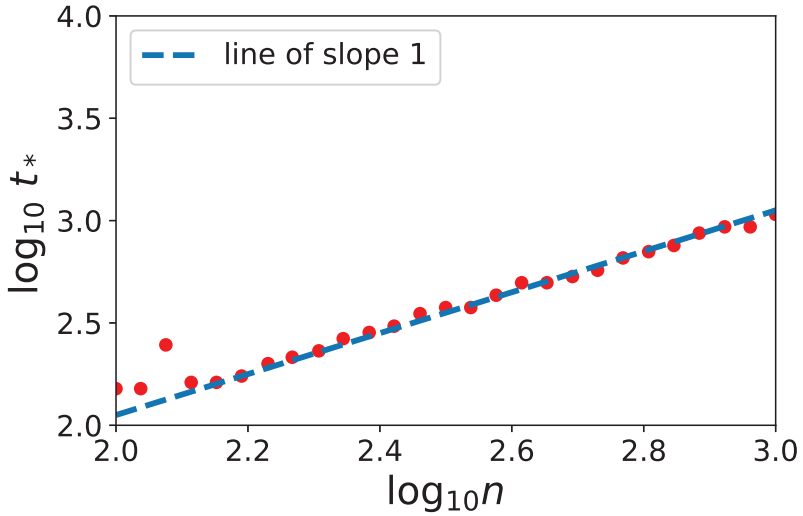


$\alpha = 3, r = 1/6 < (\alpha - 1)/(2\alpha)$

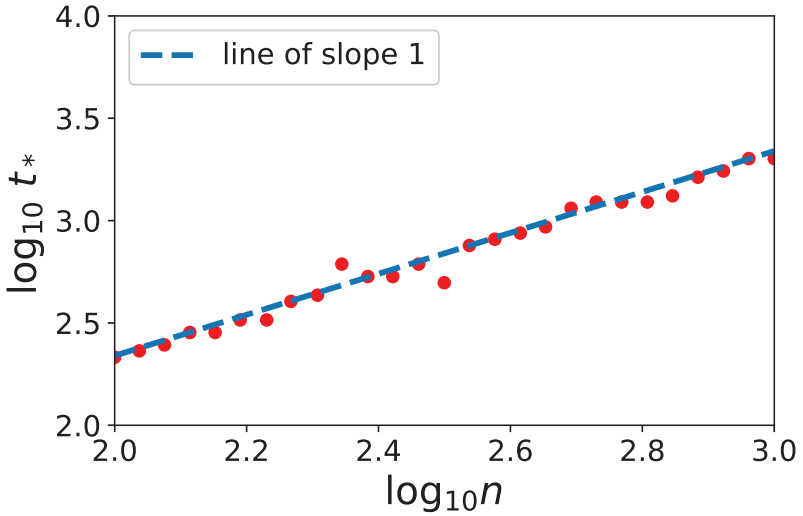


Simulations (cycling)

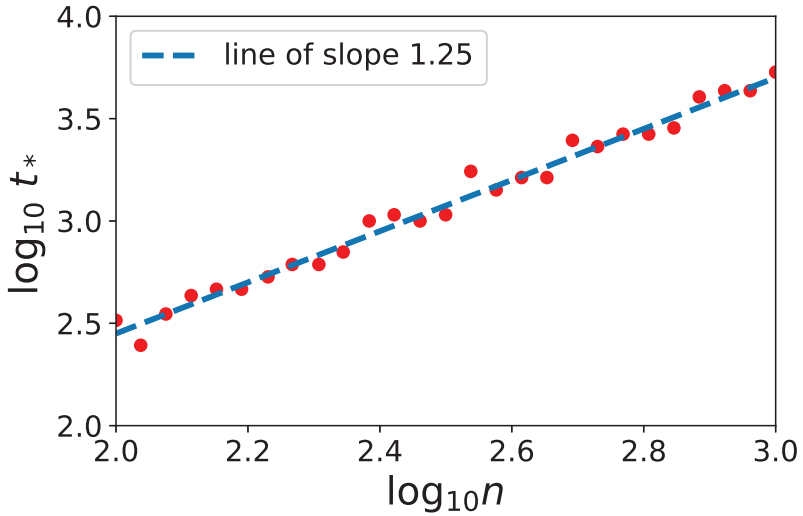
$\alpha = 3/2, r = 1/3 > (\alpha - 1)/(2\alpha)$



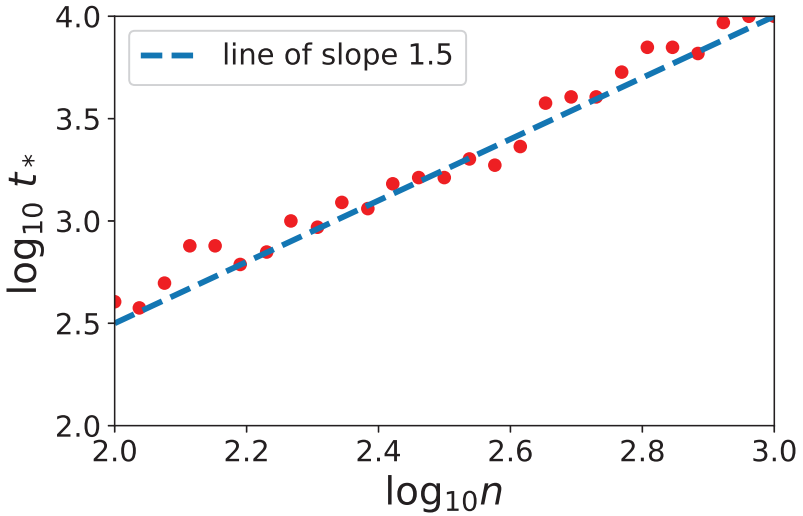
$\alpha = 4, r = 1/4 = (\alpha - 1)/(2\alpha)$



$\alpha = 5/2, r = 1/5 < (\alpha - 1)/(2\alpha)$

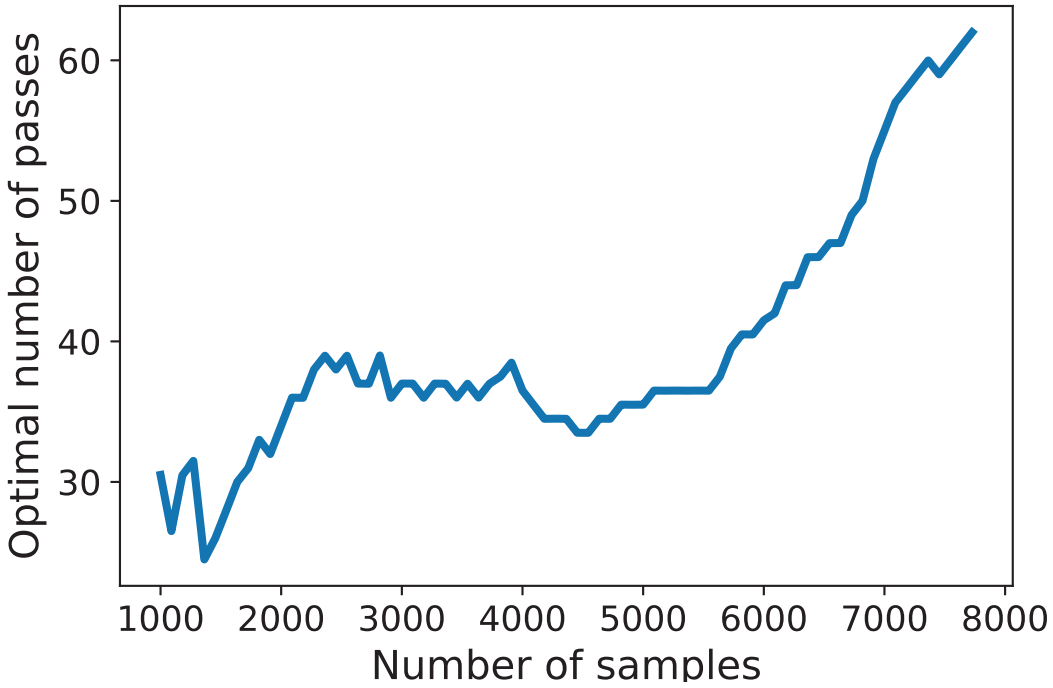


$\alpha = 3, r = 1/6 < (\alpha - 1)/(2\alpha)$



Simulations - Benchmarks

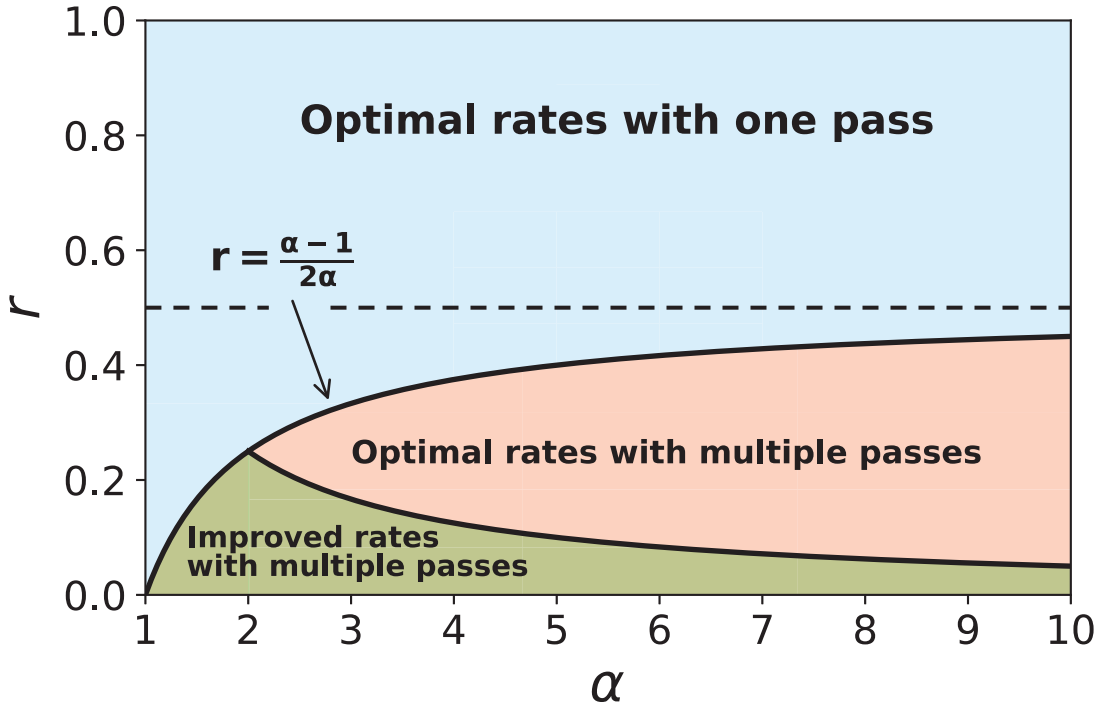
- **MNIST dataset with linear kernel**



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)

References

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Dieuleveut and F. Bach. Non-parametric Stochastic Approximation with Large Step sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. Technical Report 1707.06386, arXiv, 2017.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Why random reshuffling beats stochastic gradient descent. Technical Report 1510.08560, arXiv, 2015.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

O. Macchi. *Adaptive processing: The least mean squares approach with applications in transmission*. Wiley West Sussex, 1995.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Stochastic gradient methods with exponential convergence of testing errors. Technical Report 1712.04755, arXiv, 2017.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. Technical Report 1805.10074, arXiv, 2018.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems 29*, pages 46–54, 2016.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proc. COLT*, 2009.

A. B. Tsybakov. Optimal rates of aggregation. In *Proc. COLT*, 2003.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.