Beyond stochastic gradient descent for large-scale machine learning

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with Aymeric Dieuleveut, Nicolas Flammarion, Eric Moulines - ERNSI Workshop, 2017

Context and motivations

- **Supervised machine learning**
  - **Goal**: estimating a function $f : \mathcal{X} \to \mathcal{Y}$
  - From random observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$

Context and motivations

- **Supervised machine learning**
  - **Goal**: estimating a function $f : \mathcal{X} \to \mathcal{Y}$
  - From random observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$

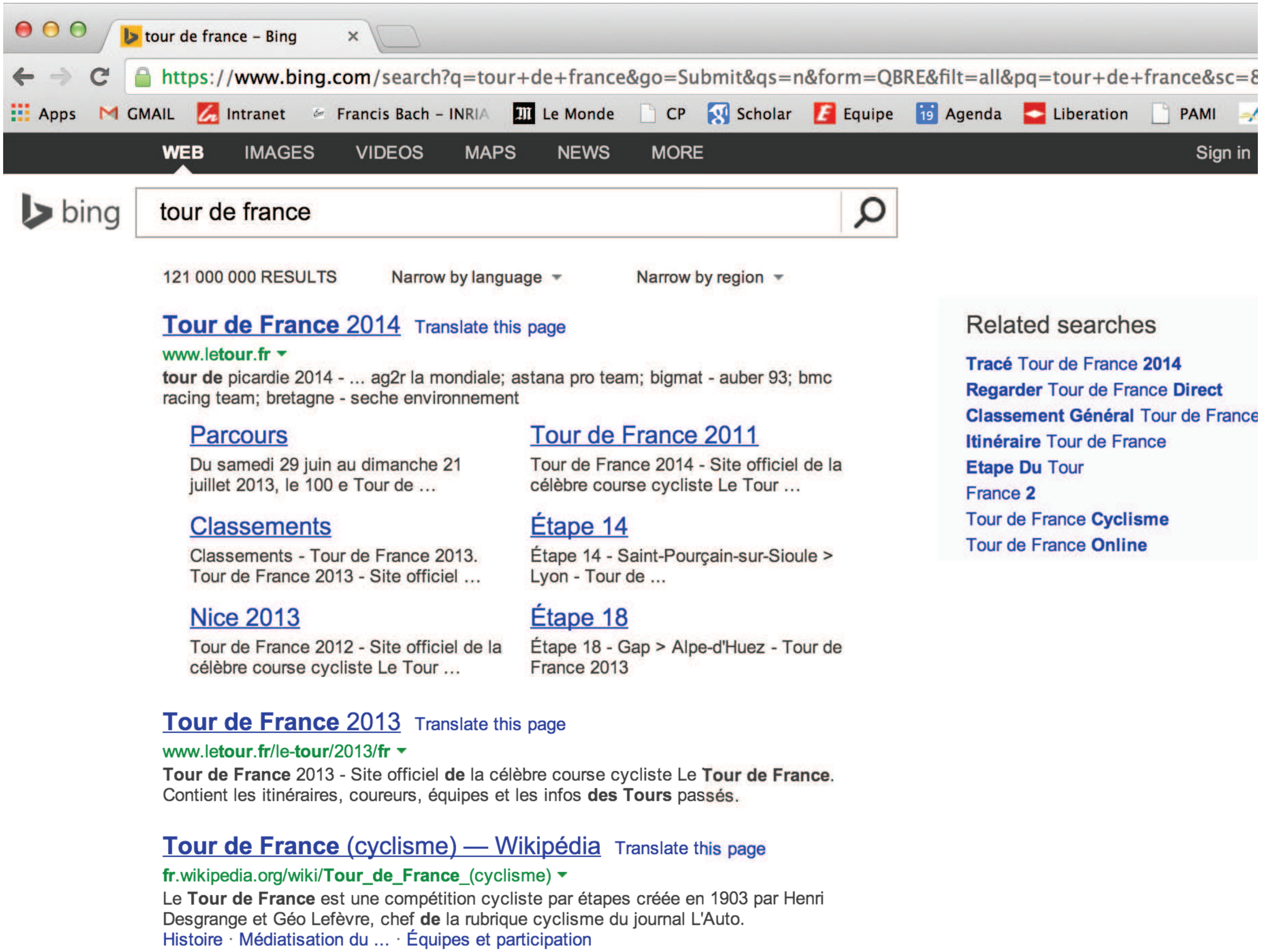
- **Specificities**
  - Data may come from anywhere
  - From strong to weak prior knowledge
  - Computational constraints
  - Between theory, algorithms and applications

Context and motivations

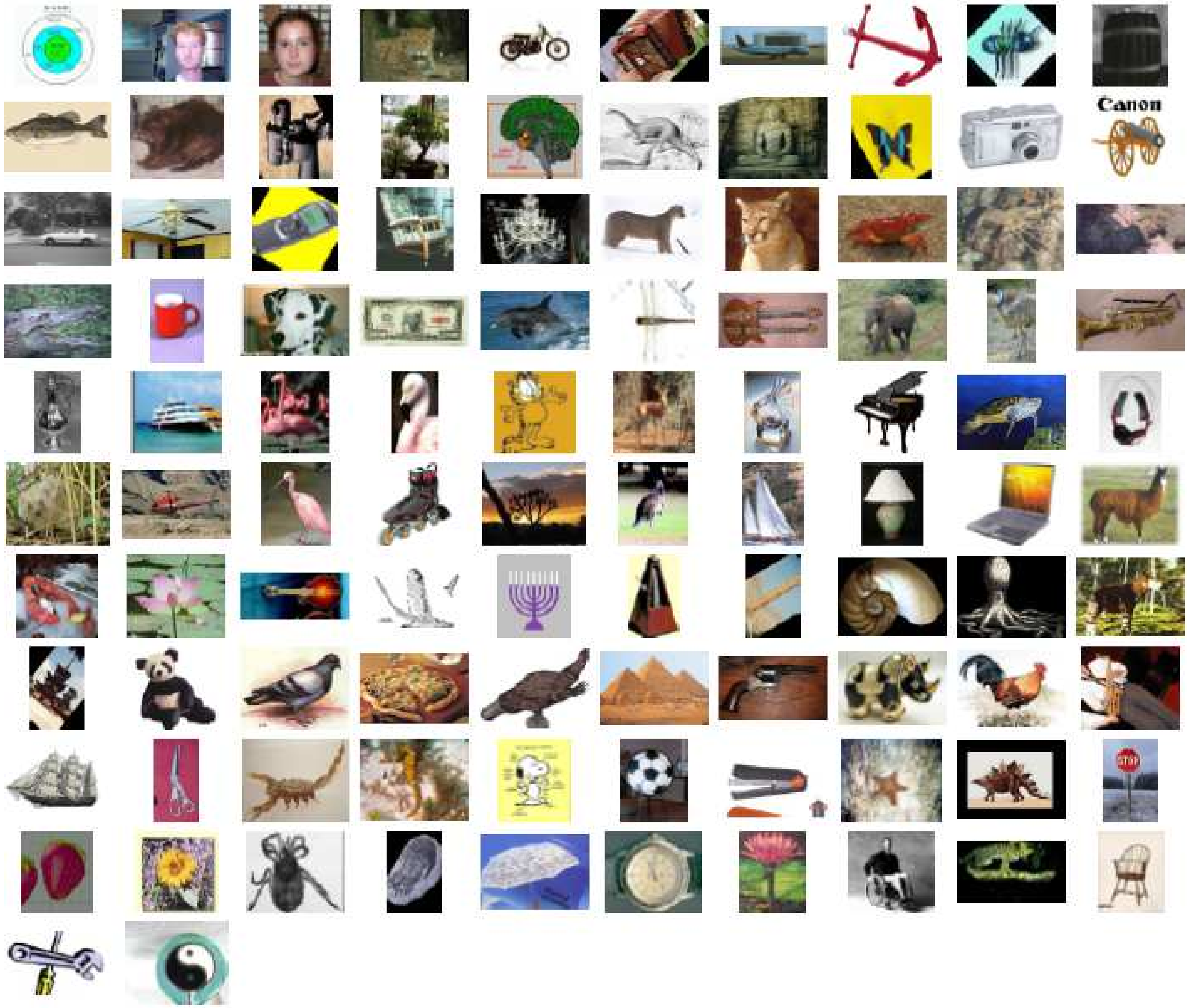
- **Large-scale machine learning**: **large p, large n**
  - p : dimension of each observation (input)
  - n : number of observations

- **Examples**: computer vision, bioinformatics, advertising

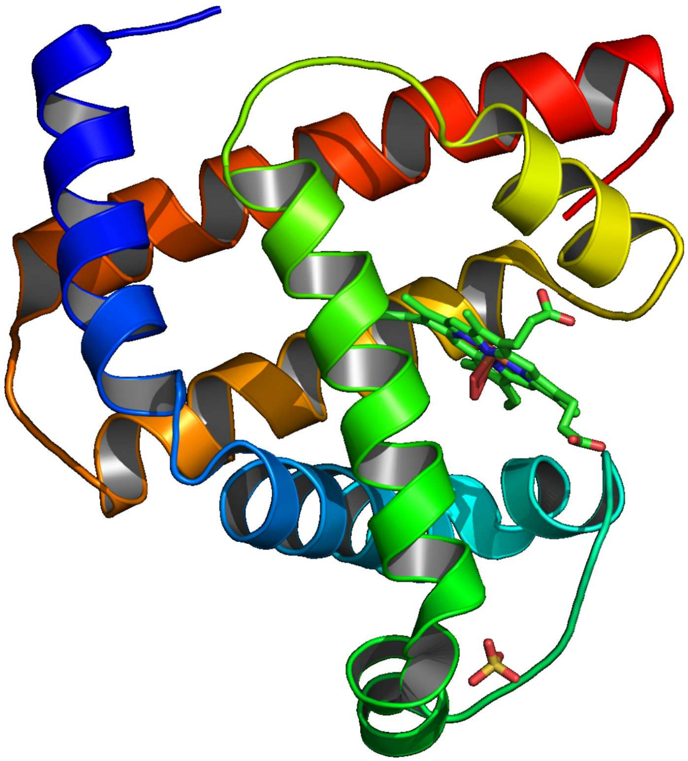
Search engines - Advertising



Visual object recognition



Bioinformatics



- **Protein**: Crucial elements of cell life
- **Massive data**: 2 millions for humans
- **Complex data**

Context and motivations

- **Large-scale machine learning**: **large p, large n**
  - p : dimension of each observation (input)
  - n : number of observations
- **Examples**: computer vision, bioinformatics, advertising
- **Ideal running-time complexity**: $O(pn)$

Context and motivations

- **Large-scale machine learning**: **large p, large n**
  - p : dimension of each observation (input)
  - n : number of observations

- **Examples**: computer vision, bioinformatics, advertising

- **Ideal running-time complexity**: $O(pn)$

- **Going back to simple methods**
  - Stochastic gradient methods (Robbins and Monro, 1951)
  - Mixing statistics and optimization

Supervised machine learning

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

- Prediction as a **linear function** $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
  - Explicit features adapted to inputs (can be learned as well)
  - Using Hilbert spaces for non-linear / non-parametric estimation

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a **linear function** $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, \langle \theta, \Phi(x_i) \rangle\big) \quad + \quad \mu \Omega(\theta)$$

convex data fitting term + regularizer

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a **linear function** $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

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

(least-squares regression)

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Prediction as a **linear function** $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

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

(logistic regression)

Supervised machine learning

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

- Prediction as a **linear function** $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$

- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, \langle \theta, \Phi(x_i) \rangle\big) \quad + \quad \mu \Omega(\theta)$$

convex data fitting term + regularizer

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$ training cost

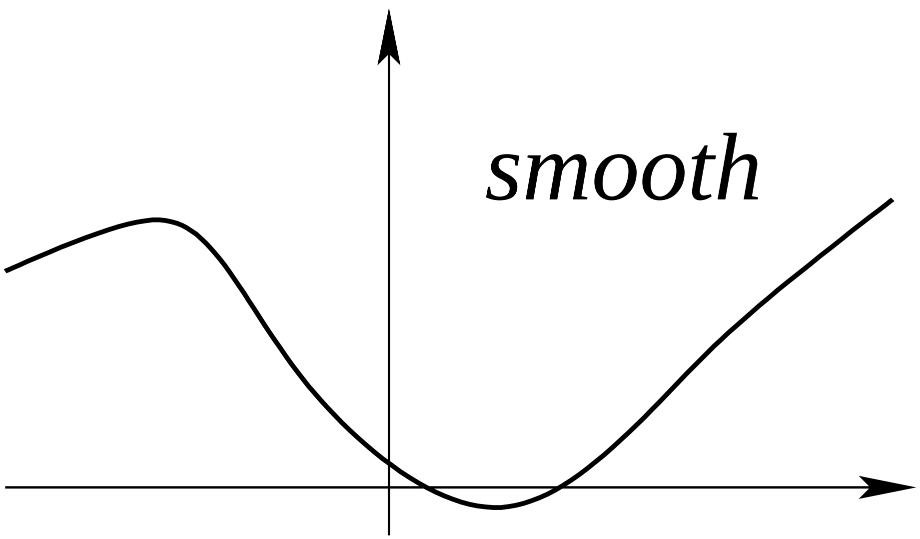
- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \langle \theta, \Phi(x) \rangle)$ testing cost

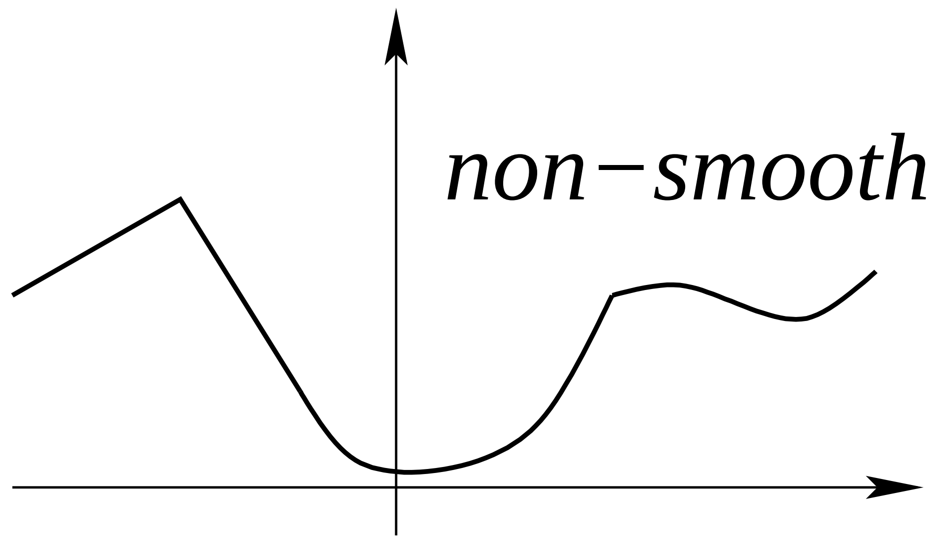
- **Two fundamental questions**: (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$

Smoothness and strong convexity

- A function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is L-smooth if and only if it is twice differentiable and

$$\forall \theta \in \mathbb{R}^p, \ \text{eigenvalues}\big[g''(\theta)\big] \leqslant L$$





Smoothness and strong convexity

- A function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is L-smooth if and only if it is twice differentiable and

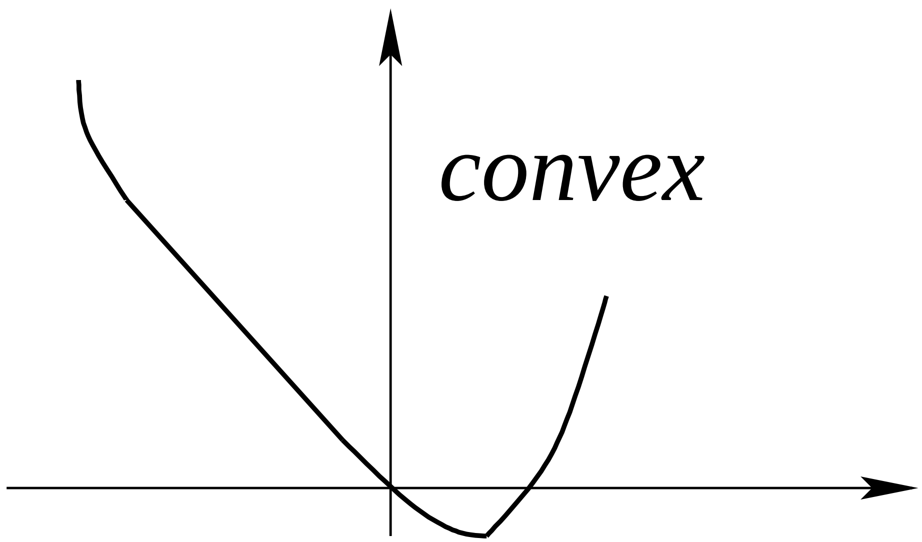
$$\forall \theta \in \mathbb{R}^p, \ \text{eigenvalues}\big[g''(\theta)\big] \leqslant L$$

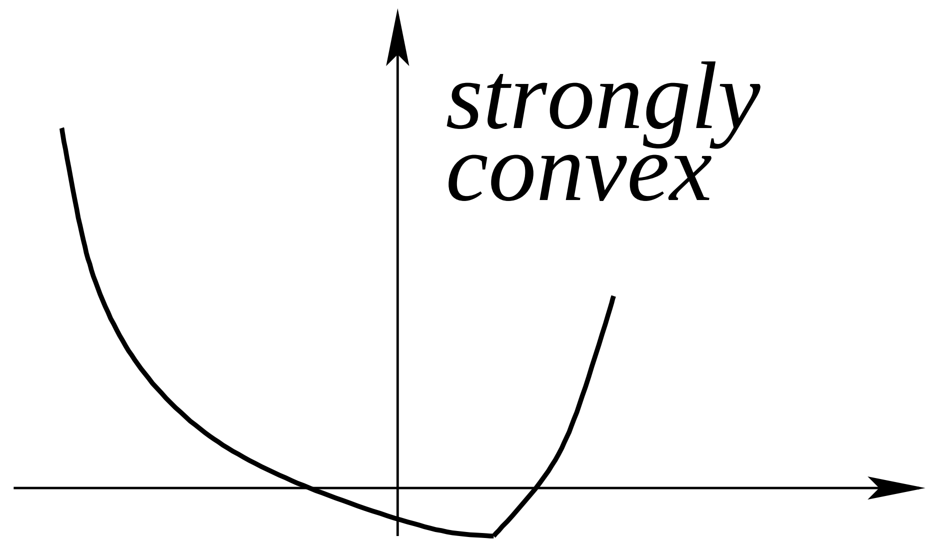
- **Machine learning**
  - with $g(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i)\rangle)$
  - Hessian $\approx$ covariance matrix $\frac{1}{n}\sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$
  - Bounded data: $\|\Phi(x)\| \leqslant R \Rightarrow L = O(R^2)$

Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, \ \text{eigenvalues}\big[g''(\theta)\big] \geqslant \mu$$

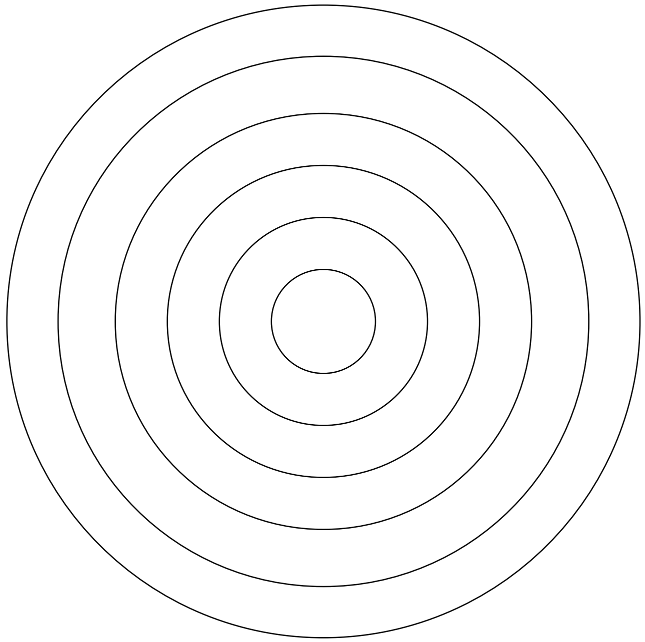




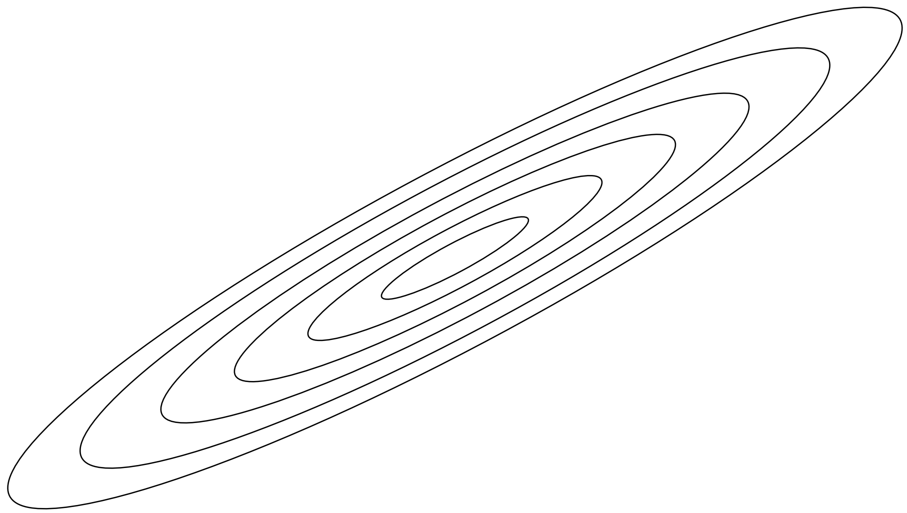
Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, \text{ eigenvalues}\big[g''(\theta)\big] \geqslant \mu$$



(large μ/L)



(small μ/L)

Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, \text{ eigenvalues}\big[g''(\theta)\big] \geqslant \mu$$

- **Machine learning**
  - with $g(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i)\rangle)$
  - Hessian $\approx$ covariance matrix $\frac{1}{n}\sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$
  - Data with invertible covariance matrix (low correlation/dimension)

Smoothness and strong convexity

- A twice differentiable function $g : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall \theta \in \mathbb{R}^p, \ \text{eigenvalues}\big[g''(\theta)\big] \geqslant \mu$$

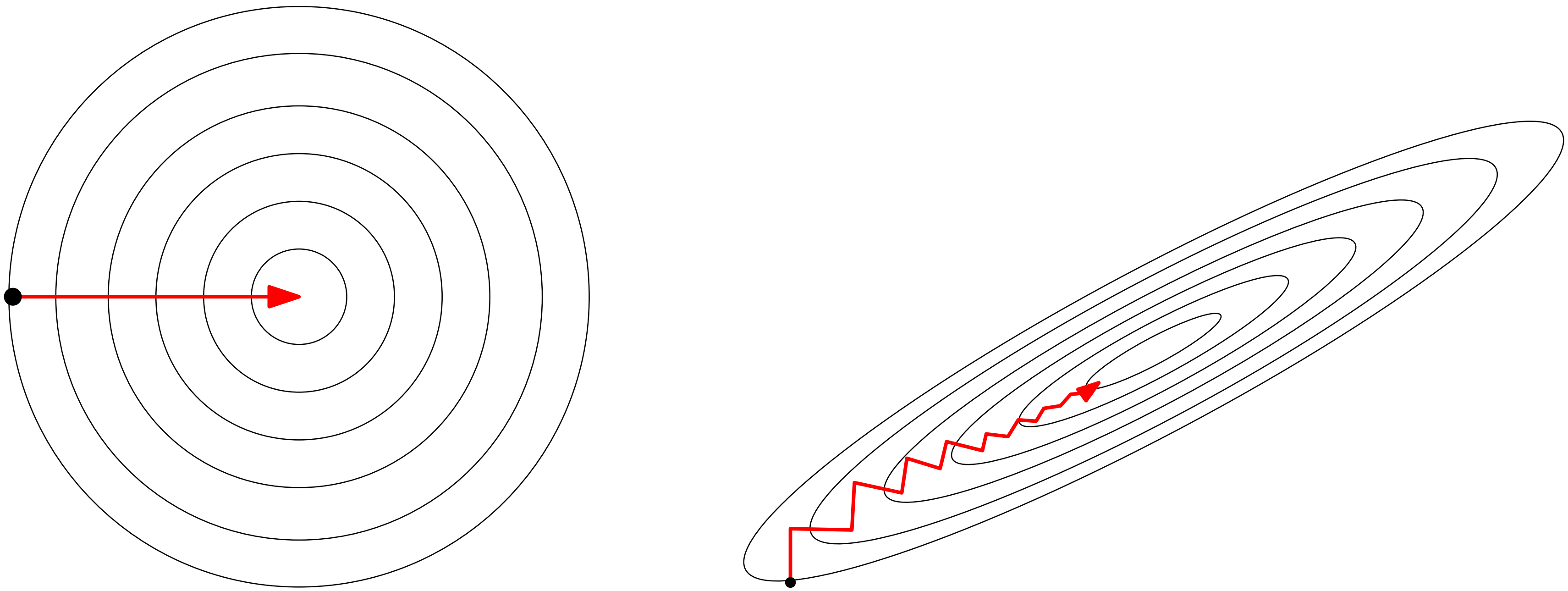
- **Machine learning**
  - with $g(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle \theta, \Phi(x_i)\rangle)$
  - Hessian $\approx$ covariance matrix $\frac{1}{n}\sum_{i=1}^n \Phi(x_i) \otimes \Phi(x_i)$
  - Data with invertible covariance matrix (low correlation/dimension)

- **Adding regularization by** $\frac{\mu}{2}\|\theta\|^2$
  - creates additional bias unless μ is small

Iterative methods for minimizing smooth functions

- **Assumption**: g convex and smooth on $\mathbb{R}^p$

- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/t)$ convergence rate for convex functions
  - $O(e^{-(\mu/L)t})$ convergence rate for strongly convex functions



Iterative methods for minimizing smooth functions

- **Assumption**: g convex and smooth on $\mathbb{R}^p$
- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/t)$ convergence rate for convex functions
  - $O(e^{-(\mu/L)t})$ convergence rate for strongly convex functions
- **Newton method**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ convergence rate

Iterative methods for minimizing smooth functions

- **Assumption**: g convex and smooth on $\mathbb{R}^p$

- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/t)$ convergence rate for convex functions
  - $O(e^{-(\mu/L)t})$ convergence rate for strongly convex functions

- **Newton method**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ convergence rate

- **Key insights from Bottou and Bousquet (2008)**
  1. In machine learning, no need to optimize below statistical error
  2. In machine learning, cost functions are averages

$\Rightarrow$ **Stochastic approximation**

Stochastic approximation

- **Goal**: Minimizing a function f defined on $\mathbb{R}^p$
  - given only unbiased estimates $f'_n(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^p$

Stochastic approximation

- **Goal**: Minimizing a function f defined on $\mathbb{R}^p$
  - given only unbiased estimates $f_n'(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^p$

- **Machine learning - statistics**
  - $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E}\, \ell(y_n, \langle \theta, \Phi(x_n) \rangle)$ = **generalization error**
  - **Loss for a single pair of observations**: $f_n(\theta) = \ell(y_n, \langle \theta, \Phi(x_n) \rangle)$
  - Expected gradient:

$$f'(\theta) = \mathbb{E} f_n'(\theta) = \mathbb{E}\left\{ \ell'(y_n, \langle \theta, \Phi(x_n) \rangle)\, \Phi(x_n) \right\}$$

- Beyond convex optimization: see, e.g., Benveniste et al. (2012)

Convex stochastic approximation

- **Key assumption**: smoothness and/or strong convexity
- **Key algorithm:** stochastic gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n\, f_n'(\theta_{n-1})}$$

  – Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\theta_k$
  – Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = C n^{-\alpha}}$

Convex stochastic approximation

- **Key assumption**: smoothness and/or strong convexity
- **Key algorithm:** stochastic gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})}$$

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\theta_k$
  - Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = Cn^{-\alpha}}$

- **Running-time** $= O(np)$
  - Single pass through the data
  - One line of code among many

Convex stochastic approximation
Existing analysis

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovski and Yudin, 1983; Agarwal et al., 2012)
  - Strongly convex: $O((\mu n)^{-1})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
  - Non-strongly convex: $O(n^{-1/2})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

Convex stochastic approximation
Existing analysis

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovski and Yudin, 1983; Agarwal et al., 2012)
  - Strongly convex: $O((\mu n)^{-1})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
  - Non-strongly convex: $O(n^{-1/2})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
  - All step sizes $\gamma_n = Cn^{-\alpha}$ with $\alpha \in (1/2, 1)$ lead to $O(n^{-1})$ for smooth strongly convex problems

Convex stochastic approximation
Existing analysis

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovski and Yudin, 1983; Agarwal et al., 2012)
  - Strongly convex: $O((\mu n)^{-1})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
  - Non-strongly convex: $O(n^{-1/2})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
  - All step sizes $\gamma_n = Cn^{-\alpha}$ with $\alpha \in (1/2, 1)$ lead to $O(n^{-1})$ for smooth strongly convex problems

- **A single algorithm for smooth problems with global convergence rate $O(1/n)$ in all situations?**

Least-mean-square algorithm

- **Least-squares**: $f(\theta) = \frac{1}{2}\mathbb{E}\big[(y_n - \langle \Phi(x_n), \theta \rangle)^2\big]$ with $\theta \in \mathbb{R}^p$
  - SGD $=$ least-mean-square algorithm (see, e.g., Macchi, 1995)
  - usually studied without averaging and decreasing step-sizes
  - with strong convexity assumption $\mathbb{E}\big[\Phi(x_n) \otimes \Phi(x_n)\big] = H \succcurlyeq \mu \cdot \mathrm{Id}$

Least-mean-square algorithm

- **Least-squares**: $f(\theta) = \frac{1}{2}\mathbb{E}\big[(y_n - \langle \Phi(x_n), \theta\rangle)^2\big]$ with $\theta \in \mathbb{R}^p$
  - SGD $=$ least-mean-square algorithm (see, e.g., Macchi, 1995)
  - usually studied without averaging and decreasing step-sizes
  - with strong convexity assumption $\mathbb{E}\big[\Phi(x_n) \otimes \Phi(x_n)\big] = H \succcurlyeq \mu \cdot \mathrm{Id}$

- **New analysis for averaging and constant step-size** $\gamma = 1/(4R^2)$
  - Assume $\|\Phi(x_n)\| \leqslant R$ and $|y_n - \langle \Phi(x_n), \theta_*\rangle| \leqslant \sigma$ almost surely
  - No assumption regarding lowest eigenvalues of H
  - Main result: $\boxed{\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leqslant \dfrac{4\sigma^2 p}{n} + \dfrac{4R^2\|\theta_0 - \theta_*\|^2}{n}}$

- **Matches statistical lower bound** (Tsybakov, 2003)
  - Fewer assumptions than existing bounds for empirical risk min.

Least-squares - Proof technique

- LMS recursion:

$$\theta_n - \theta_* = \big[I - \gamma \Phi(x_n) \otimes \Phi(x_n)\big](\theta_{n-1} - \theta_*) + \gamma\, \varepsilon_n \Phi(x_n)$$

- Simplified LMS recursion: with $H = \mathbb{E}\big[\Phi(x_n) \otimes \Phi(x_n)\big]$

$$\theta_n - \theta_* = \big[I - \gamma H\big](\theta_{n-1} - \theta_*) + \gamma\, \varepsilon_n \Phi(x_n)$$

  – Direct proof technique of Polyak and Juditsky (1992), e.g.,

$$\theta_n - \theta_* = \big[I - \gamma H\big]^n(\theta_0 - \theta_*) + \gamma \sum_{k=1}^{n} \big[I - \gamma H\big]^{n-k} \varepsilon_k \Phi(x_k)$$

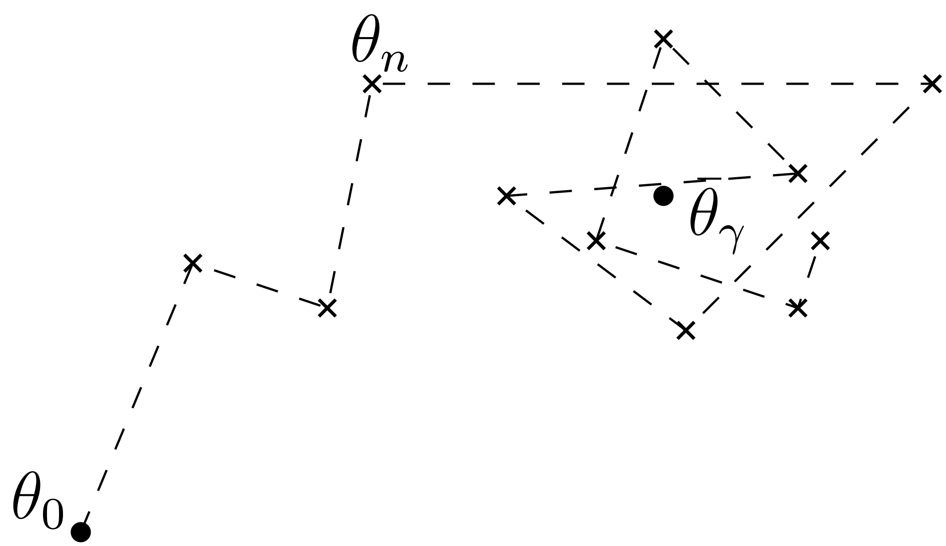
- Infinite expansion of Aguech, Moulines, and Priouret (2000) in powers of γ

Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}\big(y_n - \langle \Phi(x_n), \theta\rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle \Phi(x_n), \theta_{n-1}\rangle - y_n\big)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$



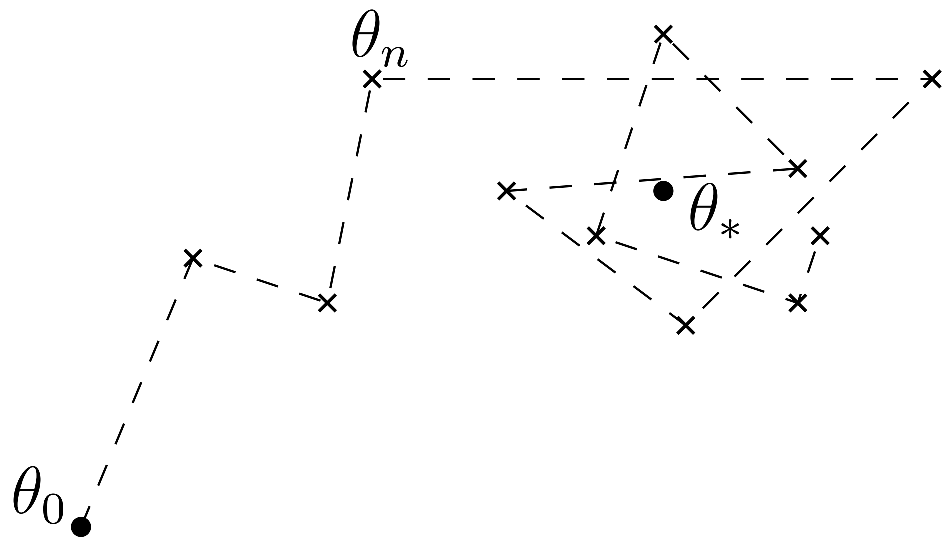
Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}\big(y_n - \langle \Phi(x_n), \theta \rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n\big)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ 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, $\bar{\theta}_\gamma = \theta_*$**



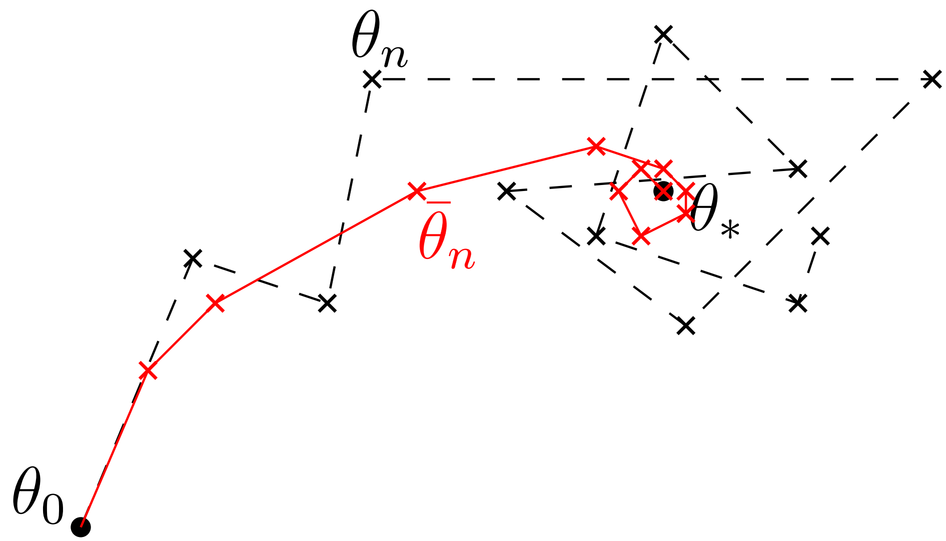
Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}\big(y_n - \langle \Phi(x_n), \theta\rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle \Phi(x_n), \theta_{n-1}\rangle - y_n\big)\Phi(x_n)$$

- The sequence $(\theta_n)_n$ 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 for $f_n(\theta) = \frac{1}{2}\big(y_n - \langle \Phi(x_n), \theta \rangle\big)^2$

$$\theta_n = \theta_{n-1} - \gamma\big(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n\big)\Phi(x_n)$$

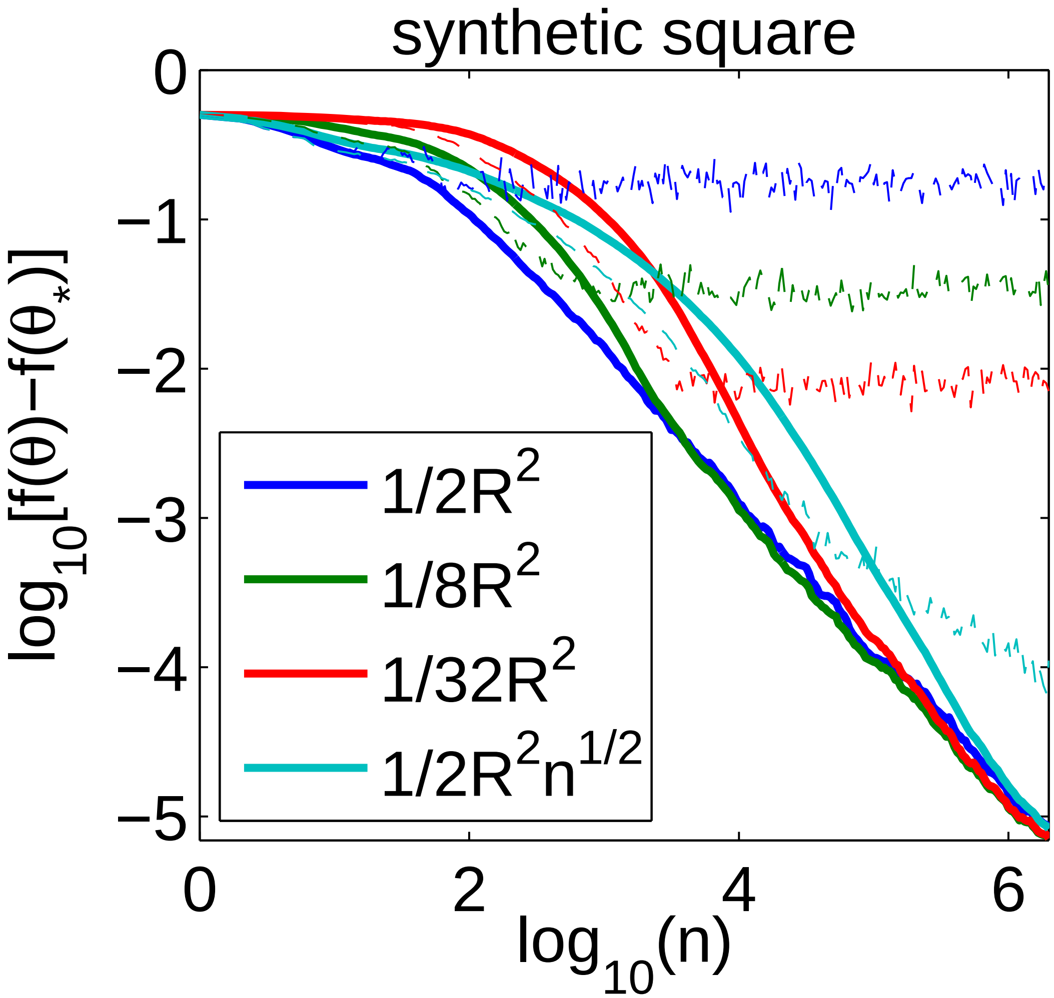
- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$

- **For least-squares, $\bar{\theta}_\gamma = \theta_*$**
  - θ_n does not converge to θ_* but oscillates around it
  - oscillations of order $\sqrt{\gamma}$

- **Ergodic theorem:**
  - Averaged iterates converge to $\bar{\theta}_\gamma = \theta_*$ at rate $O(1/n)$

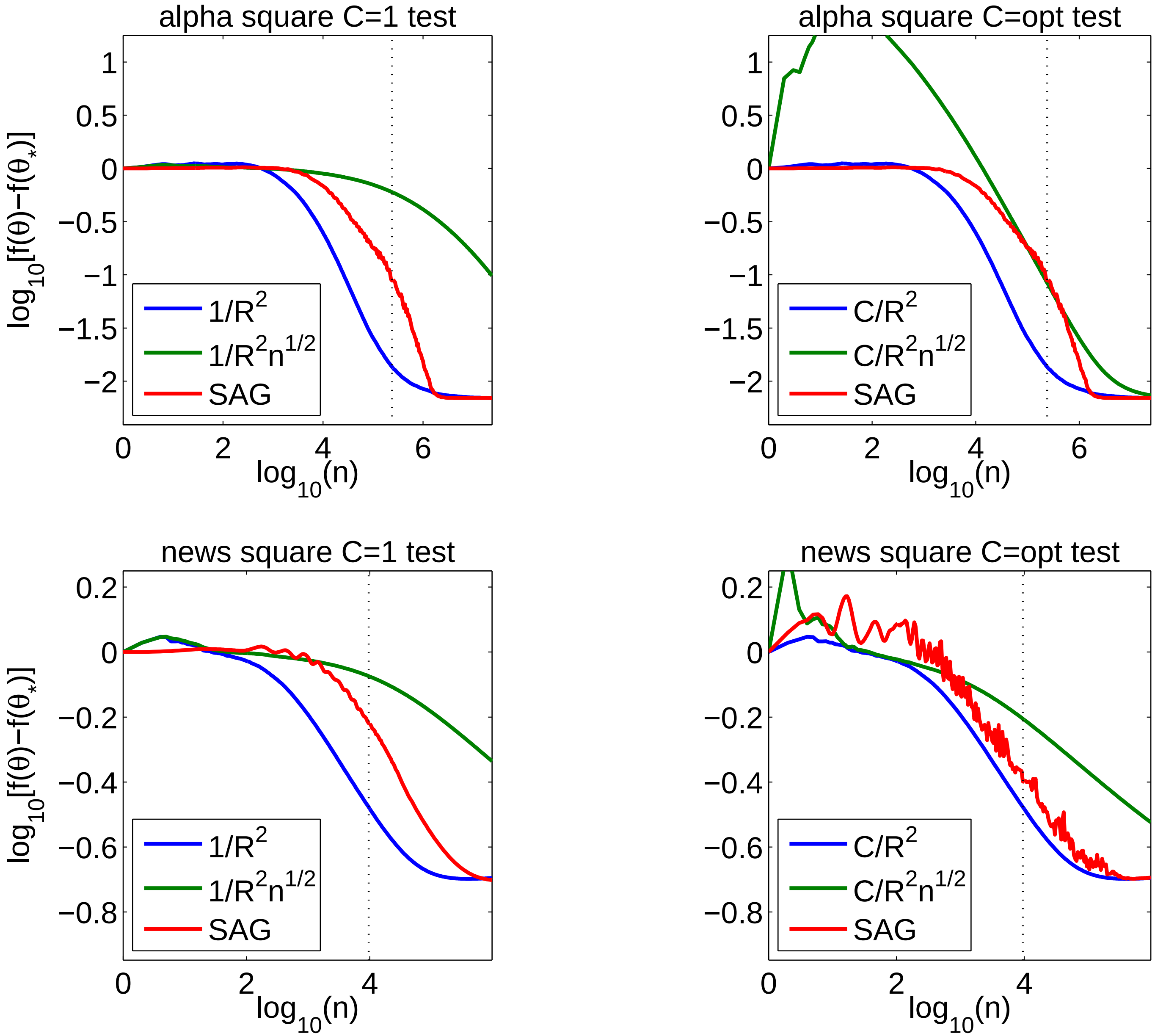
Simulations - synthetic examples

- Gaussian distributions - $p = 20$



Simulations - benchmarks

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



Isn't least-squares regression a "regression"?

Isn't least-squares regression a "regression"?

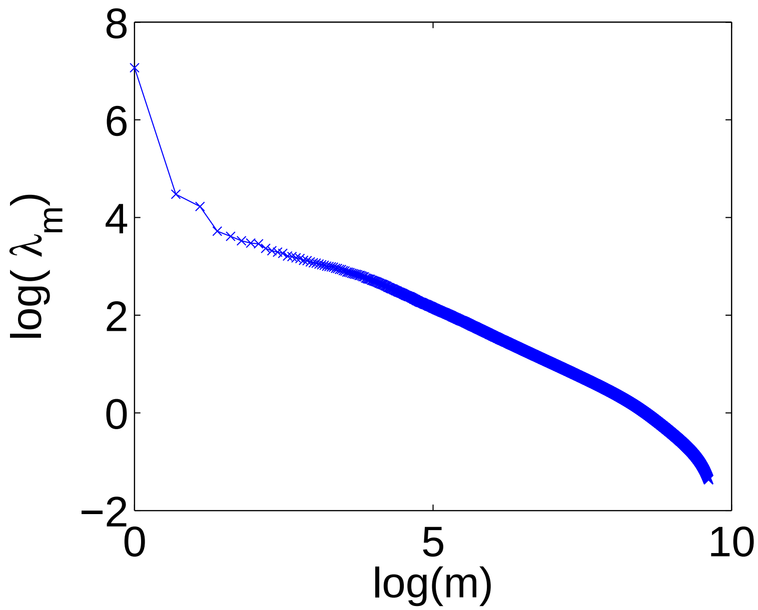
- **Least-squares regression**
  - Simpler to analyze and understand
  - Explicit relationship to bias/variance trade-offs
- **Many important loss functions are not quadratic**
  - Beyond least-squares with online Newton steps
  - Complexity of $O(p)$ per iteration with rate $O(p/n)$
  - See Bach and Moulines (2013) for details

Optimal bounds for least-squares?

- **Least-squares**: cannot beat $\sigma^2 p/n$ (Tsybakov, 2003). Really?
  - Adaptivity to simpler problems

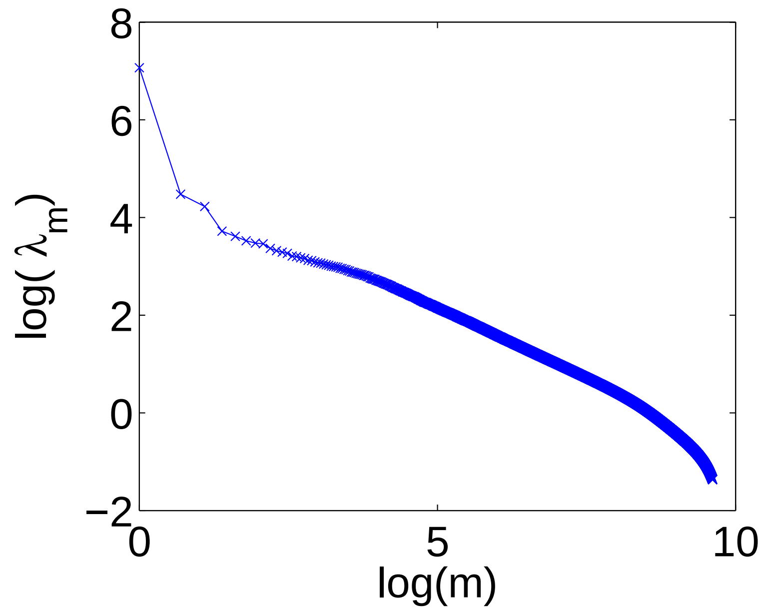
Finer assumptions (Dieuleveut and Bach, 2016)

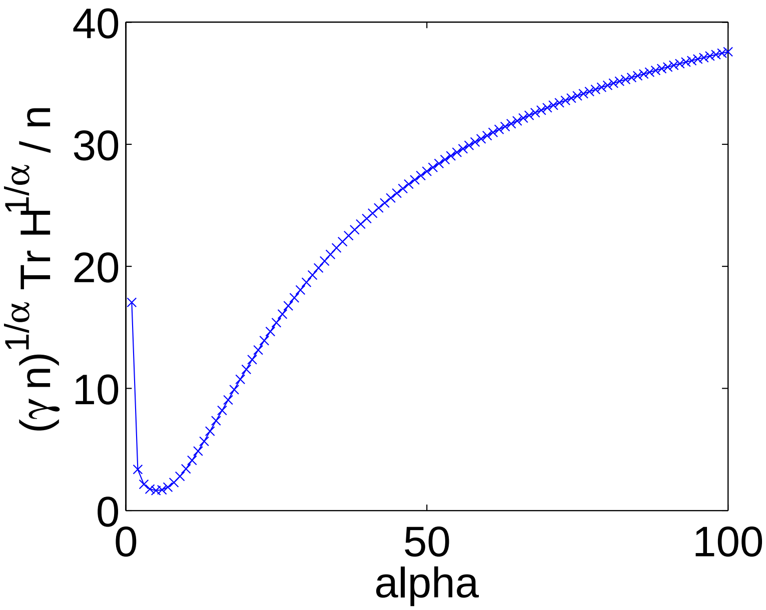
- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues λ_m less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\operatorname{tr} H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small



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 $\operatorname{tr} H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\dfrac{\sigma^2 p}{n}$ by $\dfrac{\sigma^2 (\gamma n)^{1/\alpha} \operatorname{tr} H^{1/\alpha}}{n}$





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 $\operatorname{tr} H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\dfrac{\sigma^2 p}{n}$ by $\dfrac{\sigma^2 (\gamma n)^{1/\alpha} \operatorname{tr} H^{1/\alpha}}{n}$
- **Optimal predictor**
  - Pessimistic assumption: $\|\theta_0 - \theta_*\|^2$ finite
  - Finer assumption: $\|H^{1/2-r}(\theta_0 - \theta_*)\|_2$ small
  - Replace $\dfrac{\|\theta_0 - \theta_*\|^2}{\gamma n}$ by $\dfrac{4\|H^{1/2-r}(\theta_0 - \theta_*)\|_2}{\gamma^{2r} n^{2\min\{r,1\}}}$

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 $\operatorname{tr} H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small
  - New result: replace $\dfrac{\sigma^2 p}{n}$ by $\dfrac{\sigma^2 (\gamma n)^{1/\alpha} \operatorname{tr} H^{1/\alpha}}{n}$

- **Optimal predictor**
  - Pessimistic assumption: $\|\theta_0 - \theta_*\|^2$ finite
  - Finer assumption: $\|H^{1/2-r}(\theta_0 - \theta_*)\|_2$ small
  - Replace $\dfrac{\|\theta_0 - \theta_*\|^2}{\gamma n}$ by $\dfrac{4\|H^{1/2-r}(\theta_0 - \theta_*)\|_2}{\gamma^{2r} n^{2\min\{r,1\}}}$

- **Leads to optimal rates for non-parametric regression**

Achieving optimal bias and variance terms

	Bias	Variance
Averaged gradient descent (Bach and Moulines, 2013)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n}$	$\frac{\sigma^2 p}{n}$

Achieving optimal bias and variance terms

	Bias	Variance
Averaged gradient descent (Bach and Moulines, 2013)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n}$	$\frac{\sigma^2 p}{n}$
Accelerated gradient descent (Nesterov, 1983)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n^2}$	$\sigma^2 p$

- **Acceleration is notoriously non-robust to noise** (d'Aspremont, 2008; Schmidt et al., 2011)
  - For non-structured noise, see Lan (2012)

Achieving optimal bias and variance terms

	Bias	Variance
Averaged gradient descent (Bach and Moulines, 2013)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n}$	$\frac{\sigma^2 p}{n}$
Accelerated gradient descent (Nesterov, 1983)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n^2}$	$\sigma^2 p$
"Between" averaging and acceleration (Flammarion and Bach, 2015)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n^{1+\alpha}}$	$\frac{\sigma^2 p}{n^{1-\alpha}}$

Achieving optimal bias and variance terms

	Bias	Variance
Averaged gradient descent (Bach and Moulines, 2013)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n}$	$\frac{\sigma^2 p}{n}$
Accelerated gradient descent (Nesterov, 1983)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n^2}$	$\sigma^2 p$
"Between" averaging and acceleration (Flammarion and Bach, 2015)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n^{1+\alpha}}$	$\frac{\sigma^2 p}{n^{1-\alpha}}$
Averaging and acceleration (Dieuleveut, Flammarion, and Bach, 2016)	$\frac{R^2\|\theta_0 - \theta_*\|^2}{n^2}$	$\frac{\sigma^2 p}{n}$

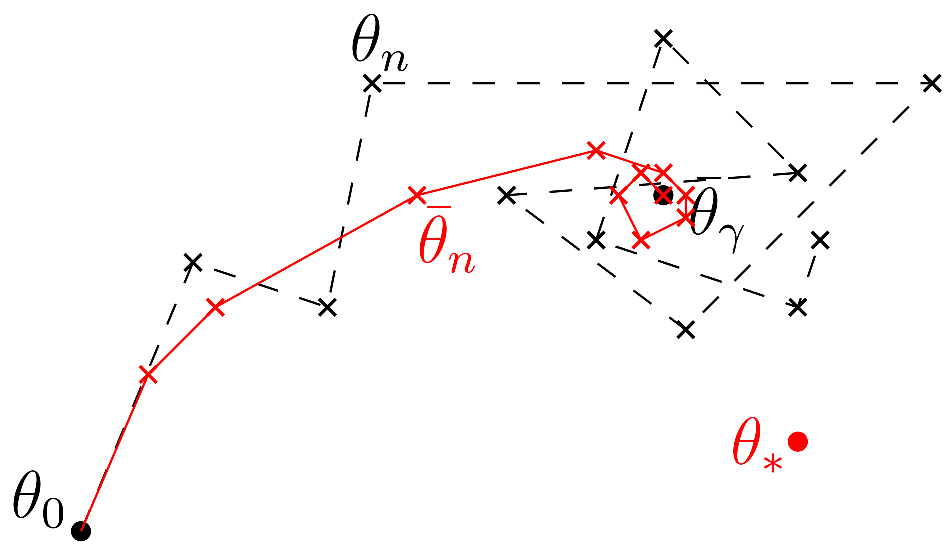
Beyond least-squares - Markov chain interpretation

- Recursion $\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1})$ also defines a Markov chain
  - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$
  - When f' is not linear, $f'(\int \theta\pi_\gamma(\mathrm{d}\theta)) \neq \int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$

Beyond least-squares - Markov chain interpretation

- Recursion $\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1})$ also defines a Markov chain
  - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$
  - When f' is not linear, $f'(\int \theta\pi_\gamma(\mathrm{d}\theta)) \neq \int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$

- **θ_n oscillates around the wrong value $\bar{\theta}_\gamma \neq \theta_*$**



Beyond least-squares - Markov chain interpretation

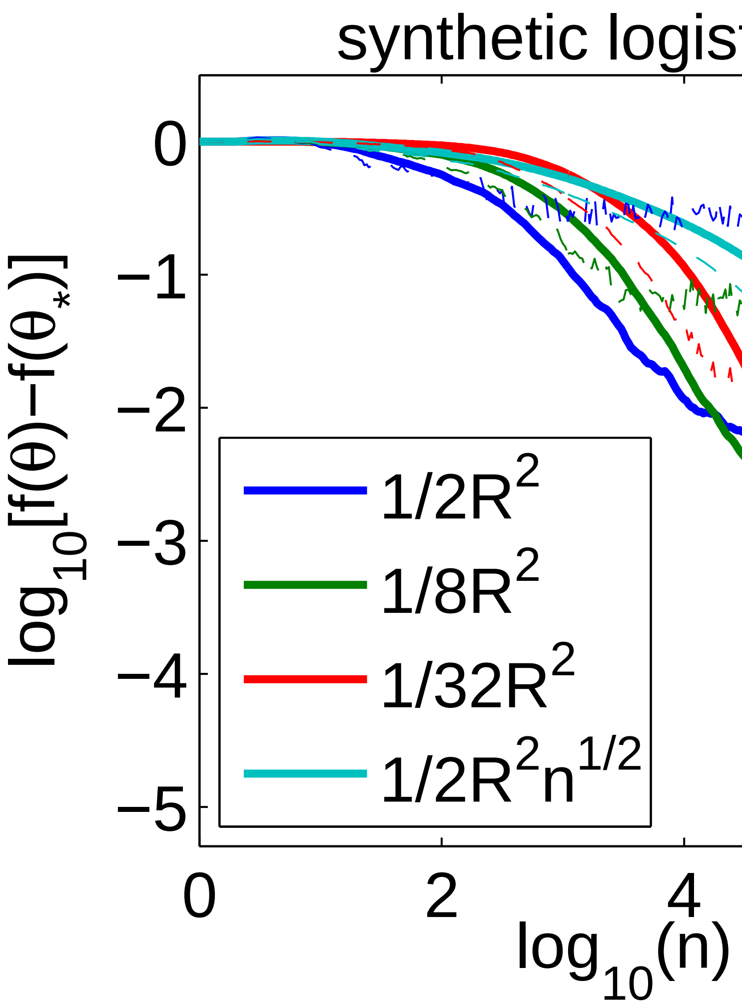
- Recursion $\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1})$ also defines a Markov chain
  - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$
  - When f' is not linear, $f'(\int \theta\pi_\gamma(\mathrm{d}\theta)) \neq \int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$

- **θ_n oscillates around the wrong value $\bar{\theta}_\gamma \neq \theta_*$**
  - moreover, $\|\theta_* - \theta_n\| = O_p(\sqrt{\gamma})$

- **Ergodic theorem**
  - averaged iterates converge to $\bar{\theta}_\gamma \neq \theta_*$ at rate $O(1/n)$
  - moreover, $\|\theta_* - \bar{\theta}_\gamma\| = O(\gamma)$ (Bach, 2014)
  - See precise analysis by Dieuleveut, Durmus, and Bach (2017)

Simulations - synthetic examples

- Gaussian distributions - $p = 20$



Restoring convergence through online Newton steps

- **Known facts**

1. Averaged SGD with $\gamma_n \propto n^{-1/2}$ leads to *robust* rate $O(n^{-1/2})$ for all convex functions
2. Averaged SGD with γ_n constant leads to *robust* rate $O(n^{-1})$ for all convex *quadratic* functions
3. Newton's method squares the error at each iteration for smooth functions
4. A single step of Newton's method is equivalent to minimizing the quadratic Taylor expansion

Restoring convergence through online Newton steps

- **Known facts**

  1. Averaged SGD with $\gamma_n \propto n^{-1/2}$ leads to *robust* rate $O(n^{-1/2})$ for all convex functions
  2. Averaged SGD with γ_n constant leads to *robust* rate $O(n^{-1})$ for all convex *quadratic* functions $\Rightarrow O(n^{-1})$
  3. Newton's method squares the error at each iteration for smooth functions $\Rightarrow O((n^{-1/2})^2)$
  4. A single step of Newton's method is equivalent to minimizing the quadratic Taylor expansion

- **Online Newton step**

  – Rate: $O((n^{-1/2})^2 + n^{-1}) = O(n^{-1})$
  – Complexity: $O(p)$ per iteration

Restoring convergence through online Newton steps

- The Newton step for $f = \mathbb{E} f_n(\theta) \stackrel{\text{def}}{=} \mathbb{E}\big[\ell(y_n, \langle\theta, \Phi(x_n)\rangle)\big]$ at $\tilde{\theta}$ is equivalent to minimizing the quadratic approximation

$$\begin{aligned} g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\ &= f(\tilde{\theta}) + \langle \mathbb{E} f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, \mathbb{E} f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\ &= \mathbb{E}\Big[f(\tilde{\theta}) + \langle f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle\Big] \end{aligned}$$

Restoring convergence through online Newton steps

- The Newton step for $f = \mathbb{E} f_n(\theta) \stackrel{\text{def}}{=} \mathbb{E}\big[\ell(y_n, \langle \theta, \Phi(x_n)\rangle)\big]$ at $\tilde{\theta}$ is equivalent to minimizing the quadratic approximation

$$\begin{aligned} g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\ &= f(\tilde{\theta}) + \langle \mathbb{E} f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, \mathbb{E} f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\ &= \mathbb{E}\Big[f(\tilde{\theta}) + \langle f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle\Big] \end{aligned}$$

- **Complexity of least-mean-square recursion for g is $O(p)$**

$$\theta_n = \theta_{n-1} - \gamma\big[f_n'(\tilde{\theta}) + f_n''(\tilde{\theta})(\theta_{n-1} - \tilde{\theta})\big]$$

  – $f_n''(\tilde{\theta}) = \ell''(y_n, \langle \tilde{\theta}, \Phi(x_n)\rangle)\Phi(x_n) \otimes \Phi(x_n)$ has rank one
  – New online Newton step without computing/inverting Hessians

Choice of support point for online Newton step

- **Two-stage procedure**

(1) Run $n/2$ iterations of averaged SGD to obtain $\tilde{\theta}$
(2) Run $n/2$ iterations of averaged constant step-size LMS

  - Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)
  - Provable convergence rate of $O(p/n)$ for logistic regression
  - Additional assumptions but no strong convexity

Choice of support point for online Newton step

- **Two-stage procedure**

(1) Run $n/2$ iterations of averaged SGD to obtain $\tilde{\theta}$
(2) Run $n/2$ iterations of averaged constant step-size LMS

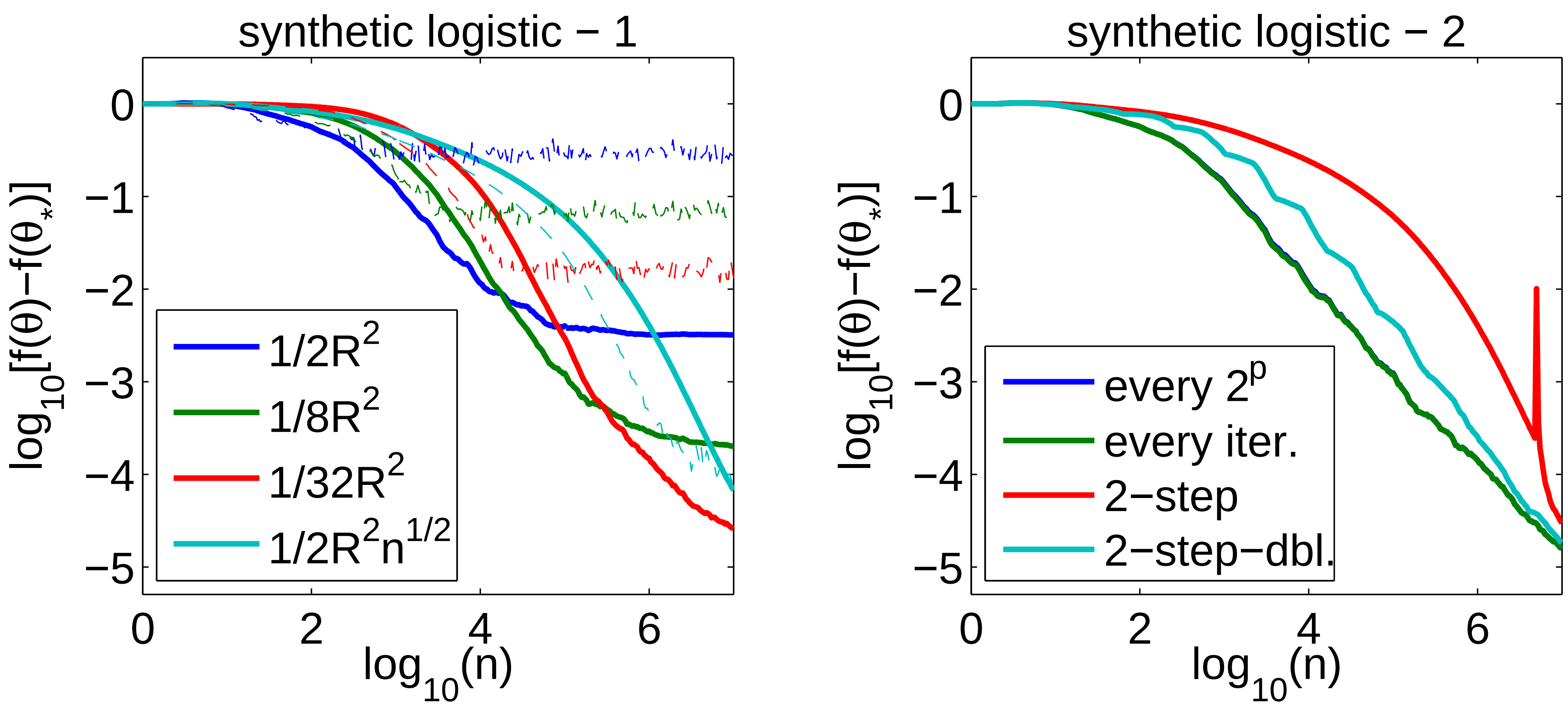
  - Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)
  - Provable convergence rate of $O(p/n)$ for logistic regression
  - Additional assumptions but no strong convexity

- **Update at each iteration using the current averaged iterate**

  - Recursion: $\boxed{\theta_n = \theta_{n-1} - \gamma\big[f'_n(\bar{\theta}_{n-1}) + f''_n(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1})\big]}$
  - No provable convergence rate (yet) but best practical behavior
  - Note (dis)similarity with regular SGD: $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$

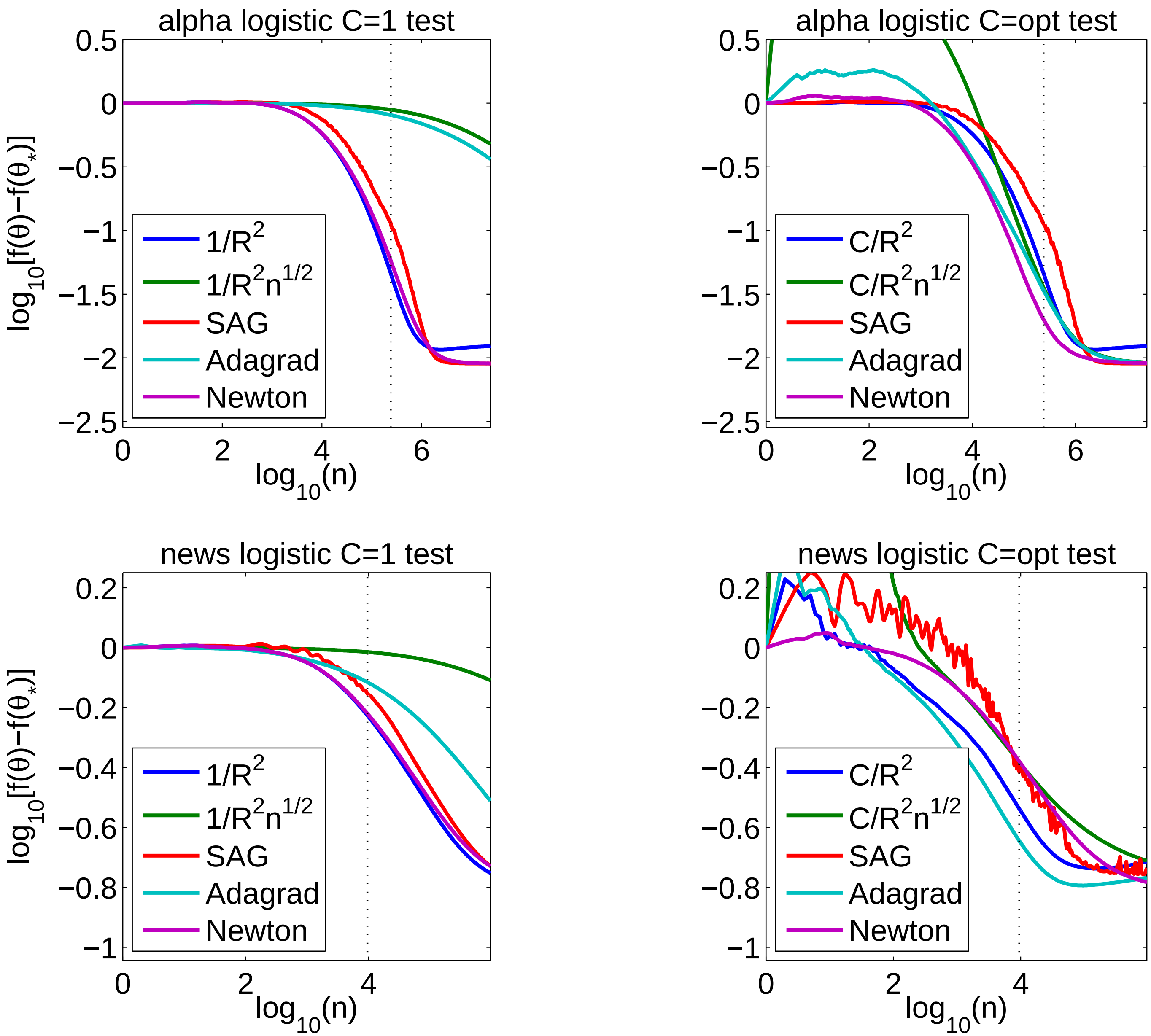
Simulations - synthetic examples

- Gaussian distributions - $p = 20$



Simulations - benchmarks

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



Conclusions

- **Constant-step-size averaged stochastic gradient descent**
  - Reaches convergence rate $O(1/n)$ in all regimes
  - Improves on the $O(1/\sqrt{n})$ lower-bound of non-smooth problems
  - Efficient online Newton step for non-quadratic problems
  - Robustness to step-size selection and adaptivity

Conclusions

- **Constant-step-size averaged stochastic gradient descent**
  - Reaches convergence rate $O(1/n)$ in all regimes
  - Improves on the $O(1/\sqrt{n})$ lower-bound of non-smooth problems
  - Efficient online Newton step for non-quadratic problems
  - Robustness to step-size selection and adaptivity

- **Extensions and future work**
  - Going beyond a single pass (Le Roux, Schmidt, and Bach, 2012; Defazio, Bach, and Lacoste-Julien, 2014)
  - Non-differentiable regularization (Flammarion and Bach, 2017)
  - Kernels and nonparametric estimation (Dieuleveut and Bach, 2016)
  - Parallelization
  - Non-convex problems

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 58(5):3235–3249, 2012.

R. Aguech, E. Moulines, and P. Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control and Optimization*, 39(3):872–899, 2000.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

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

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer Publishing Company, Incorporated, 2012.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Adv. NIPS*, 2008.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183, 2008.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. Technical Report 1602.05419, arXiv, 2016.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. Technical report, HAL - to appear, 2017.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. *arXiv preprint arXiv:1504.01577*, 2015.

N. Flammarion and F. Bach. Stochastic composite least-squares regression with convergence rate $O(1/n)$. In *Proc. COLT*, 2017.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A): 365–397, 2012.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

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

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley & Sons, 1983.

Y. Nesterov. A method for solving a convex programming problem with rate of convergence $O(1/k^2)$. *Soviet Math. Doklady*, 269(3):543–547, 1983.

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

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates for inexact proximal-gradient method. In *Adv. NIPS*, 2011.

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

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge Univ. press, 2000.