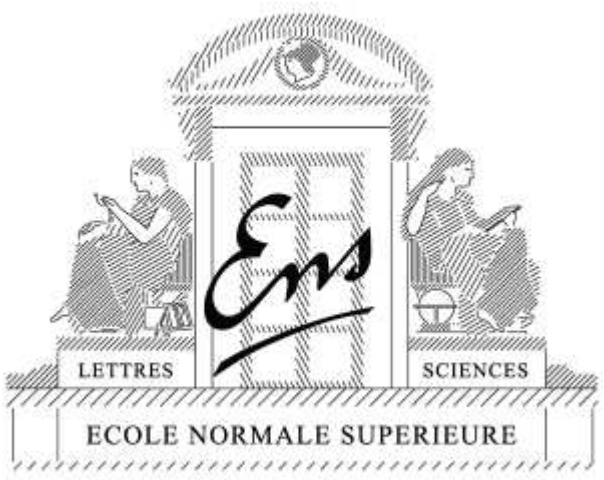
Machine learning and convex optimization

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France



Joint work with **Eric Moulines** (*Telecom Paristech*)

January 2012

Outline

- **Machine learning and optimization**
  - Motivation: large-scale machine learning problems
  - Traditional statistical analysis
  - Classical methods/concepts for convex optimization
  - Statistics relevant without algorithms?
- **Analysis of stochastic approximation algorithms**
  - Stochastic gradient and averaging
  - Strongly convex vs. non-strongly convex
  - Statistics simpler with algorithms?
- **Conclusion and open problems**

Context

- **Large-scale machine learning**: **large p, large n, large k**
  - n: number of observations
  - p: size of each observation
  - k: number of tasks
- **Examples**: computer vision, bioinformatics
- **Ideal running-time complexity**: $O(pn + kn)$
- **Statistics vs. optimization**
  - Generalization error $=$ approximation error $+$ estimation error
  - Optimization error is crucial in practice (Bottou and Bousquet, 2008)
  - Going back to simple methods

Supervised learning

- Data: n observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, **i.i.d.**
- Vector space $\mathcal{F}$ of prediction functions θ
- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta(x_i)) + \mu \Omega(\theta) \quad \textbf{or} \quad \min_{\theta \in \mathcal{F},\ \Omega(\theta) \leqslant D^2} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta(x_i))$$

  - Convex loss ℓ, convex regularizer Ω

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta(x_i))$
- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta(x))$
- **Two fundamental questions**: (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$

Analysis of empirical risk minimization

- **Approximation and estimation errors**: $\mathcal{C} = \{\theta \in \mathcal{F}, \Omega(\theta) \leqslant D^2\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathcal{F}} f(\theta) \; = \; \Big[f(\hat{\theta}) - \min_{\theta \in \mathcal{C}} f(\theta) \Big] + \Big[\min_{\theta \in \mathcal{C}} f(\theta) - \min_{\theta \in \mathcal{F}} f(\theta) \Big]$$

1. **Uniform deviation bounds**, with $\hat{\theta} \in \arg\min_{\theta \in \mathcal{C}} \hat{f}(\theta)$:

$$f(\hat{\theta}) - \min_{\theta \in \mathcal{C}} f(\theta) \;\; \leqslant \;\; 2 \sup_{\theta \in \mathcal{C}} |\hat{f}(\theta) - f(\theta)|$$

   – Typically slow rate $O(\frac{1}{\sqrt{n}})$

2. **More refined concentration results** with faster rates

Slow rate for supervised learning

- **Assumptions** (f is the expected risk, $\hat{f}$ the empirical risk)
  - $\mathcal{F}$ Hilbert space, and $\Omega(\theta) = \|\theta\|^2$
  - "Linear" predictors: $\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\| \leqslant B$ a.s.
  - L-Lipschitz loss, i.e., f and $\hat{f}$ are LB-Lipschitz on $\mathcal{C} = \{\|\theta\| \leqslant D\}$

- With probability greater than $1 - \delta$

$$\sup_{\theta \in \mathcal{C}} |\hat{f}(\theta) - f(\theta)| \leqslant \frac{BLD}{\sqrt{n}} \left[2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- Expectated estimation error: $\mathbb{E}\big[\sup_{\theta \in \mathcal{C}} |\hat{f}(\theta) - f(\theta)| \big] \leqslant \dfrac{2BLD}{\sqrt{n}}$

- Using Rademacher averages (see, e.g., Boucheron et al., 2005)

- **Lipschitz functions $\Rightarrow$ slow rate**

Fast rate for supervised learning

- **Assumptions** (f is the expected risk, $\hat{f}$ the empirical risk)
  - $\mathcal{F}$ Hilbert space, and $\Omega(\theta) = \|\theta\|^2$
  - "Linear" predictors: $\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\| \leqslant B$ a.s.
  - L-Lipschitz loss, i.e., f and $\hat{f}$ are LB-Lipschitz on $\mathcal{C} = \{\|\theta\| \leqslant D\}$.
  - The risk f is μ-strongly convex

- For any $a > 0$, with probability greater than $1 - \delta$, for all $\theta \in \mathcal{C}$,

$$f(\theta) - \min_{\eta \in \mathcal{C}} f(\eta) \leqslant (1+a)(\hat{f}(\theta) - \min_{\eta \in \mathcal{C}} \hat{f}(\eta)) + \frac{8(1+\frac{1}{a})L^2B^2(32 + \log\frac{1}{\delta})}{\mu n}$$

- Results from Sridharan, Srebro, and Shalev-Shwartz (2008)
  - More subtle concentration results
  - see also Boucheron and Massart (2011) and references therein

- **Strongly convex functions $\Rightarrow$ fast rate**

Outline

- **Machine learning and optimization**
  - Motivation: large-scale machine learning problems
  - Traditional statistical analysis
  - Classical methods/concepts for convex optimization
  - Statistics relevant without algorithms?
- **Analysis of stochastic approximation algorithms**
  - Stochastic gradient and averaging
  - Strongly convex vs. non-strongly convex
  - Statistics simpler with algorithms?
- **Conclusion and open problems**

Complexity results in convex optimization

- **Assumption**: f convex on $\mathcal{H}$ (Hilbert space or $\mathbb{R}^p$)
- **Classical generic algorithms**
  - (sub)gradient descent
  - Accelerated gradient descent
  - Newton method
- **Key additional properties of f**
  - Lipschitz continuity, smoothness or strong convexity
- **Key insight from Bottou and Bousquet (2008)**
  - In machine learning, no need to optimize below estimation error
- **Key reference**: Nesterov (2004)

Smoothness/convexity assumptions

- **Bounded gradients of f (Lipschitz-continuity)**: the function f if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathcal{H}, \|\theta\| \leqslant D \Rightarrow \|f'(\theta)\| \leqslant B$$

Smoothness/convexity assumptions

- **Bounded gradients of f (Lipschitz-continuity)**: the function f if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathcal{H}, \|\theta\| \leqslant D \Rightarrow \|f'(\theta)\| \leqslant B$$

- **Smoothness of f**: the function f is convex, differentiable with L-Lipschitz-continuous gradient f':

$$\forall \theta_1, \theta_2 \in \mathcal{H}, \quad \|f'(\theta_1) - f'(\theta_2)\| \leqslant L\|\theta_1 - \theta_2\|$$

Smoothness/convexity assumptions

- **Bounded gradients of f (Lipschitz-continuity)**: the function f if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathcal{H}, \|\theta\| \leqslant D \Rightarrow \|f'(\theta)\| \leqslant B$$

- **Smoothness of f**: the function f is convex, differentiable with L-Lipschitz-continuous gradient f':

$$\forall \theta_1, \theta_2 \in \mathcal{H}, \quad \|f'(\theta_1) - f'(\theta_2)\| \leqslant L\|\theta_1 - \theta_2\|$$

- **Strong convexity of f**: The function f is strongly convex with respect to the norm $\|\cdot\|$, with convexity constant $\mu > 0$:

$$\forall \theta_1, \theta_2 \in \mathcal{H}, \ f(\theta_1) \geqslant f(\theta_2) + \langle f'(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2$$

Subgradient descent

- **Assumptions**
  - f convex and B-Lipschitz-continuous on $\{\|\theta\| \leqslant D\}$
- **Algorithm**:

$$\theta_n = \Pi_D \left(\theta_{n-1} - \frac{D}{B\sqrt{2n}} f'(\theta_{n-1}) \right)$$

- **Bound**:

$$f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k \right) - f(\theta^*) \leqslant \frac{2DB}{\sqrt{n}}$$

- Three-line proof
- Minimax convergence rate

Subgradient descent - strong convexity

- **Assumptions**
  - f convex and B-Lipschitz-continuous on $\{\|\theta\| \leqslant D\}$
  - f μ-strongly convex
- **Algorithm**:

$$\theta_n = \Pi_D\left(\theta_{n-1} - \frac{1}{\mu n} f'(\theta_{n-1})\right)$$

- **Bound**:

$$f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta^*) \leqslant \frac{B^2 \log(n+1)}{\mu n}$$

- Three-line proof
- Minimax convergence rate

(smooth) gradient descent

- **Assumptions**
  - f convex with L-Lipschitz-continuous gradient
  - Minimum attained at θ^*
- **Algorithm**:

$$\theta_n = \theta_{n-1} - \frac{1}{L} f'(\theta_{n-1})$$

- **Bound**:

$$f(\theta_n) - f(\theta^*) \leqslant \frac{2L\|\theta_0 - \theta^*\|^2}{n+4}$$

- Four-line proof
- Not minimax convergence rate

(smooth) gradient descent - strong convexity

- **Assumptions**
  - f convex with L-Lipschitz-continuous gradient
  - f μ-strongly convex
- **Algorithm**:

$$\theta_n = \theta_{n-1} - \frac{1}{L} f'(\theta_{n-1})$$

- **Bound**:

$$f(\theta_n) - f(\theta^*) \leqslant \frac{L\|\theta_0 - \theta^*\|^2}{2}(1 - \mu/L)^n$$

- Four-line proof
- **Adaptivity of gradient descent to problem difficulty**
- Line search

Accelerated gradient methods (Nesterov, 1983)

- **Assumptions**
  - f convex with L-Lipschitz-cont. gradient , min. attained at θ^*
- **Algorithm**:

$$\begin{aligned} \theta_n &= \eta_{n-1} - \frac{1}{L} f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \frac{n-1}{n+2}(\theta_n - \theta_{n-1}) \end{aligned}$$

- **Bound**:

$$f(\theta_n) - f(\theta^*) \leqslant \frac{2L\|\theta_0 - \theta^*\|^2}{(n+1)^2}$$

- Ten-line proof (see, e.g., Schmidt, Le Roux, and Bach, 2011)
- Extension to strongly convex functions

Optimization for sparsity-inducing norms
(see Bach, Jenatton, Mairal, and Obozinski, 2011)

- Gradient descent as a **proximal method** (differentiable functions)
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^p} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2} \|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \theta_t - \frac{1}{L} \nabla f(\theta_t)$

Optimization for sparsity-inducing norms
(see Bach, Jenatton, Mairal, and Obozinski, 2011)

- Gradient descent as a **proximal method** (differentiable functions)
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^p} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2}\|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \theta_t - \frac{1}{L}\nabla f(\theta_t)$
- Problems of the form: $\boxed{\min_{\theta \in \mathbb{R}^p} f(\theta) + \mu\Omega(\theta)}$
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^p} L(\theta_t) + (\theta - \theta_t)^\top \nabla L(\theta_t) + \mu\Omega(\theta) + \frac{L}{2}\|\theta - \theta_t\|_2^2$
  - $\Omega(\theta) = \|\theta\|_1 \Rightarrow$ **Thresholded gradient descent**
- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

Outline

- **Machine learning and optimization**
  - Motivation: large-scale machine learning problems
  - Traditional statistical analysis
  - Classical methods/concepts for convex optimization
  - Statistics relevant without algorithms?
- **Analysis of stochastic approximation algorithms**
  - Stochastic gradient and averaging
  - Strongly convex vs. non-strongly convex
  - Statistics simpler with algorithms?
- **Conclusion and open problems**

Stochastic approximation

- **Goal**: Minimizing a function f defined on a Hilbert space $\mathcal{H}$
  - given only unbiased estimates $f'_n(\theta_n)$ of its (sub)gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathcal{H}$

- **Stochastic approximation**
  - Observation of $f'_n(\theta_n) = f'(\theta_n) + \varepsilon_n$
  - $\varepsilon_n =$ additive noise (typically i.i.d.)
  - May only observe a function which is positively correlated to $f'(\theta_n)$

Stochastic approximation

- **Goal**: Minimizing a function f defined on a Hilbert space $\mathcal{H}$
  - given only unbiased estimates $f'_n(\theta_n)$ of its (sub)gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathcal{H}$

- **Stochastic approximation**
  - Observation of $f'_n(\theta_n) = f'(\theta_n) + \varepsilon_n$
  - $\varepsilon_n =$ additive noise (typically i.i.d.)
  - May only observe a function which is positively correlated to $f'(\theta_n)$

- **Machine learning - statistics**
  - $f_n(\theta) = \ell(\theta, z_n)$ where z_n is an i.i.d. sequence
  - $f(\theta) = \mathbb{E} f_n(\theta) =$ generalization error of predictor θ
  - Typically $f_n(\theta) = \frac{1}{2}(\langle x_n, \theta \rangle - y_n)^2$ or $\log[1 + \exp(-y_n \langle x_n, \theta \rangle)]$, for $x_n \in \mathcal{H}$ and $y_n \in \{-1, 1\}$.

Online vs. batch learning

- **Goal**: minimize $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$ **generalization error** of θ
  - Supervised learning: $\ell(\theta, z)$ of the form $\ell(\langle x, \theta \rangle, y)$ with $z = (x, y)$
- **Batch learning**
  - Finite set of observations: $z_1, \ldots, z_n$
  - Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\theta, z_i)$
  - Estimator $\hat{\theta} =$ Minimizer of $\hat{f}(\theta)$ over a certain class Θ
  - Generalization bound using uniform concentration results

Online vs. batch learning

- **Goal**: minimize $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$ **generalization error** of θ
  - Supervised learning: $\ell(\theta, z)$ of the form $\ell(\langle x, \theta \rangle, y)$ with $z = (x, y)$
- **Batch learning**
  - Finite set of observations: $z_1, \ldots, z_N$
  - Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\theta, z_k)$
  - Estimator $\hat{\theta} =$ Minimizer of $\hat{f}(\theta)$ over a certain class Θ
  - Generalization bound using uniform concentration results
- **Online learning**
  - Update $\hat{\theta}_n$ after each new (potentially adversarial) observation z_n
  - Cumulative loss: $\frac{1}{n} \sum_{k=1}^{n} \ell(\hat{\theta}_{k-1}, z_k)$
  - Online to batch through averaging (Cesa-Bianchi et al., 2004)

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex

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

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex

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

- **Desirable practical behavior**
  - Applicable (at least) to least-squares and logistic regression
  - Robustness to (potentially unknown) constants (L,B,μ)
  - Adaptivity to difficulty of the problem (e.g., strong convexity)

Convex stochastic approximation
Related work

- **Machine learning/optimization**
  - Known minimax rates of convergence (Nemirovski and Yudin, 1983; Agarwal et al., 2010)
    - Strongly convex: $O(n^{-1})$
    - Non-strongly convex: $O(n^{-1/2})$
  - Achieved with and/or without averaging (up to log terms)
  - Non-asymptotic analysis (high-probability bounds)
  - Online setting and regret bounds
  - Bottou and Le Cun (2005); Bottou and Bousquet (2008); Hazan et al. (2007); Shalev-Shwartz and Srebro (2008); Shalev-Shwartz et al. (2007, 2009); Xiao (2010); Duchi and Singer (2009)
  - Nesterov and Vial (2008); Nemirovski et al. (2009)

Convex stochastic approximation
Related work

- **Stochastic approximation**
  - Asymptotic analysis
  - Non convex case with strong convexity around the optimum
  - $\gamma_n = Cn^{-\alpha}$ with $\alpha = 1$ is not robust to the choice of C
  - $\alpha \in (1/2, 1)$ is robust with averaging
  - Broadie et al. (2009); Kushner and Yin (2003); Kul′chitskiĭ and Mozgovoĭ (1991); Polyak and Juditsky (1992); Ruppert (1988); Fabian (1968)

Problem set-up - General assumptions

- **Unbiased gradient estimates**: Let $(\mathcal{F}_n)_{n\geqslant 0}$ be an increasing family of σ-fields. θ_0 is $\mathcal{F}_0$-measurable, and for each $\theta \in \mathcal{H}$, the random variable $f_n'(\theta)$ is square-integrable, $\mathcal{F}_n$-measurable and

$$\forall \theta \in \mathcal{H}, \quad \forall n \geqslant 1, \quad \mathbb{E}(f_n'(\theta)|\mathcal{F}_{n-1}) = f'(\theta), \ \text{w.p.1}$$

- **Variance of estimates**: There exists $\sigma^2 \geqslant 0$ such that for all $n \geqslant 1$, $\mathbb{E}(\|f_n'(\theta^*)\|^2|\mathcal{F}_{n-1}) \leqslant \sigma^2, \ \text{w.p.1}$, where θ^* is a global minimizer of f

- Specificity of machine learning
  - Full function $\theta \mapsto f_n(\theta) = \ell(\theta, z_n)$ is observed
  - Beyond i.i.d. assumptions

Problem set-up - Smoothness/convexity assumptions

- **Bounded gradients of f_n:** For each $n \geqslant 1$, almost surely, the function f_n if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathcal{H},\ \forall n > 0,\ \|\theta\| \leqslant D \Rightarrow \|f_n'(\theta)\| \leqslant B$$

Problem set-up - Smoothness/convexity assumptions

- **Bounded gradients of f_n:** For each $n \geqslant 1$, almost surely, the function f_n if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathcal{H},\ \forall n > 0,\ \|\theta\| \leqslant D \Rightarrow \|f_n'(\theta)\| \leqslant B$$

- **Smoothness of f_n:** For each $n \geqslant 1$, the function f_n is a.s. convex, differentiable with L-Lipschitz-continuous gradient f_n':

$$\forall n \geqslant 1,\ \forall \theta_1, \theta_2 \in \mathcal{H},\quad \|f_n'(\theta_1) - f_n'(\theta_2)\| \leqslant L\|\theta_1 - \theta_2\|\,, \quad \text{w.p.1}$$

Problem set-up - Smoothness/convexity assumptions

- **Bounded gradients of f_n:** For each $n \geqslant 1$, almost surely, the function f_n if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathcal{H},\ \forall n > 0,\ \|\theta\| \leqslant D \Rightarrow \|f_n'(\theta)\| \leqslant B$$

- **Smoothness of f_n:** For each $n \geqslant 1$, the function f_n is a.s. convex, differentiable with L-Lipschitz-continuous gradient f_n':

$$\forall n \geqslant 1,\ \forall \theta_1, \theta_2 \in \mathcal{H},\quad \|f_n'(\theta_1) - f_n'(\theta_2)\| \leqslant L\|\theta_1 - \theta_2\|\,, \quad \text{w.p.1}$$

- **Strong convexity of f:** The function f is strongly convex with respect to the norm $\|\cdot\|$, with convexity constant $\mu > 0$:

$$\forall \theta_1, \theta_2 \in \mathcal{H},\ f(\theta_1) \geqslant f(\theta_2) + \langle f'(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2$$

Summary of new results

- Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old: $O(n^{-1})$ rate achieved without averaging for $\alpha = 1$
  - New: $O(n^{-1})$ rate achieved with averaging for $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants

Summary of new results

- Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$

- **Strongly convex smooth objective functions**
  - Old: $O(n^{-1})$ rate achieved without averaging for $\alpha = 1$
  - New: $O(n^{-1})$ rate achieved with averaging for $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants

- **Non-strongly convex smooth objective functions**
  - Old: $O(n^{-1/2})$ rate achieved with averaging for $\alpha = 1/2$
  - New: $O(\max\{n^{1/2-3\alpha/2}, n^{-\alpha/2}, n^{\alpha-1}\})$ rate achieved without averaging for $\alpha \in [1/3, 1]$,

- **Take-home message**
  - Use $\alpha = 1/2$ with averaging to be adaptive to strong convexity

Strongly convex functions
Stochastic gradient descent

- **Assumptions**: f μ-strongly convex, f_n' L-Lipschitz
- Notation: $\varphi_\beta(t) = \begin{cases} \frac{t^\beta - 1}{\beta} & \text{if } \beta \neq 0, \\ \log t & \text{if } \beta = 0. \end{cases}$
  - $\beta \mapsto \varphi_\beta(t)$ continuous $\forall t > 0$,
  - $\beta > 0 \Rightarrow \varphi_\beta(t) < \frac{t^\beta}{\beta}$ and $\beta < 0 \Rightarrow \varphi_\beta(t) < \frac{1}{-\beta}$

Strongly convex functions
Stochastic gradient descent

- **Assumptions**: f μ-strongly convex, f_n' L-Lipschitz
- Notation: $\varphi_\beta(t) = \begin{cases} \frac{t^\beta - 1}{\beta} & \text{if } \beta \neq 0, \\ \log t & \text{if } \beta = 0. \end{cases}$
  - $\beta \mapsto \varphi_\beta(t)$ continuous $\forall t > 0$,
  - $\beta > 0 \Rightarrow \varphi_\beta(t) < \frac{t^\beta}{\beta}$ and $\beta < 0 \Rightarrow \varphi_\beta(t) < \frac{1}{-\beta}$
- **Bound on** $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$, if $\gamma_n = Cn^{-\alpha}$:

$$\begin{cases} 2\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)+\frac{4C\sigma^2}{\mu n^\alpha}, & \text{if } \alpha \in [0,1] \\ \frac{\exp(2L^2C^2)}{n^{\mu C}}\left(\delta_0+\frac{\sigma^2}{L^2}\right) + 2\sigma^2C^2\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}}, & \text{if } \alpha = 1 \end{cases}$$

Strongly convex functions
Stochastic gradient descent

- **Bound on** $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$, if $\gamma_n = Cn^{-\alpha}$:

$$\begin{cases} 2\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)+\frac{4C\sigma^2}{\mu n^\alpha}, & \text{if } \alpha\in[0,1] \\ \frac{\exp(2L^2C^2)}{n^{\mu C}}\left(\delta_0+\frac{\sigma^2}{L^2}\right)+2\sigma^2C^2\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}}, & \text{if } \alpha=1 \end{cases}$$

- **Related work**: Kul′chitskiĭ and Mozgovoĭ (1991); Broadie et al. (2009); Nesterov and Vial (2008); Nemirovski et al. (2009)

- **Sketch of proof**

  1. Derive deterministic recursion

$$\delta_n \leqslant (1-2\mu\gamma_n+2L^2\gamma_n^2)\delta_{n-1}+2\sigma^2\gamma_n^2$$

  2. Mimic SA proof techniques in a non-asymptotic way

Strongly convex functions
Stochastic gradient descent

- **Bound on** $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$, if $\gamma_n = Cn^{-\alpha}$:

$$\begin{cases} 2\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)+\frac{4C\sigma^2}{\mu n^\alpha}, & \text{if } \alpha \in [0,1] \\ \frac{\exp(2L^2C^2)}{n^{\mu C}}\left(\delta_0+\frac{\sigma^2}{L^2}\right)+2\sigma^2C^2\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}}, & \text{if } \alpha = 1 \end{cases}$$

- **Discussion**
  - Forgetting initial conditions sub-exponentially fast
  - Bound on function values using smoothness
  - Tightness for quadratic functions
  - Valid for $\alpha \in [0,1]$
  - Non robust behavior for $\alpha = 1$
  - Setting C too large
  - Minimax rate for $\alpha = 1$ (but with good constant C)

Strongly convex functions
Stochastic gradient descent - Bounded gradients

- **Assumptions**: f μ-strongly convex, f_n B-Lipschitz
- Requires to stay in compact set (e.g., using orthogonal projections)
- **Bound on** $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$:

$$\delta_n \leqslant \begin{cases} \left(\delta_0 + B^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{2}n^{1-\alpha}\right) + \frac{2B^2C^2}{\mu n^\alpha}, & \text{if } \alpha \in [0,1] \\ \delta_0 n^{-\mu C} + 2B^2C^2 n^{-\mu C}\varphi_{\mu C-1}(n), & \text{if } \alpha = 1 \end{cases}$$

- **Related work**:
  - Nemirovski et al. (2009); Shalev-Shwartz et al. (2007)
- No explosive multiplicative factors

Strongly convex functions - Averaging

- **Assumptions**: f μ-strongly convex, f_n' L-Lipschitz, f_n'' M-Lipschitz
  - Noise variance: $\mathbb{E}(f_n'(\theta^*) \otimes f_n'(\theta^*) | \mathcal{F}_{n-1}) \preccurlyeq \Sigma$
- **Bound on** $(\mathbb{E}\|\bar{\theta}_n - \theta^*\|^2)^{1/2}$, if $\gamma_n = Cn^{-\alpha}$:

$$\frac{\left[\operatorname{tr} f''(\theta^*)^{-1} \Sigma f''(\theta^*)^{-1}\right]^{1/2}}{\sqrt{n}}$$

$$+ \frac{6\sigma}{\mu C^{1/2}} \frac{1}{n^{1-\alpha/2}} + \frac{MC\tau^2}{2\mu^{3/2}} (1 + (\mu C)^{1/2}) \frac{\varphi_{1-\alpha}(n)}{n}$$

$$+ \frac{4LC^{1/2}}{\mu} \frac{\varphi_{1-\alpha}(n)^{1/2}}{n} + \frac{8A}{n\mu^{1/2}} \left(\frac{1}{C} + L\right) \left(\delta_0 + \frac{\sigma^2}{L^2}\right)^{1/2}$$

$$+ \frac{5MC^{1/2}\tau}{2n\mu} A \exp\left(24L^4C^4\right) \left(\delta_0 + \frac{\mu \mathbb{E}\left[\|\theta_0 - \theta^*\|^4\right]}{20C\tau^2} + 2\tau^2 C^3 \mu + 8\tau^2 C^2\right)$$

Strongly convex functions - Averaging

- **Sketch of proof**:
  - Following Polyak and Juditsky (1992)
  - Write recursion as $f'_n(\theta_{n-1}) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n)$ and notice
    - $f'_n(\theta_{n-1}) \approx f'_n(\theta^*) + f''(\theta^*)(\theta_{n-1} - \theta^*)$,
    - $f'_n(\theta^*)$ has zero mean and behaves like an i.i.d. sequence, and
    - $\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\theta_{k-1} - \theta_k)$ turns out to be negligible owing to a summation by parts
  - This implies that $\bar{\theta}_n - \theta^*$ behaves like $-\frac{1}{n}\sum_{k=1}^n f''(\theta^*)^{-1} f'_k(\theta^*)$.

- **Discussion**:
  - Forgetting initial conditions only polynomially
  - Asymptotically leading term - "optimal" and independent of (γ_n)
  - No need to invert Hessian!
  - Relationship with prior work on online learning - case $\alpha = 1$

Non-strongly convex functions
Stochastic gradient descent

- **Assumptions**: f convex, f_n' L-Lipschitz
- **Bound on** $\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]$, if $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [1/2, 1]$:

$$\frac{1}{C}\left(\delta_0 + \frac{\sigma^2}{L^2}\right) \exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right) \frac{1 + 4L^{3/2}C^{3/2}}{\min\{\varphi_{1-\alpha}(n), \varphi_{\alpha/2}(n)\}}.$$

- **Discussion**
  - New result
  - Cases $\alpha = 1$ and $\alpha = 1/2$
  - Change of behavior at $\alpha = 2/3$
  - Conjecture: optimal rates (for stochastic gradient descent)

Non-strongly convex functions
Stochastic gradient descent - Bounded gradients

- **Assumptions**: f convex, f_n B-Lipschitz, f_n' L-Lipschitz

- **Bound on** $\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]$, if $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [1/3, 1]$:

$$\begin{cases} \left(\delta_0 + B^2C^2\varphi_{1-2\alpha}(n)\right) \frac{1+4L^{1/2}C^{1/2}}{C\min\{\varphi_{1-\alpha}(n),\varphi_{\alpha/2}(n)\}}, & \text{if } \alpha \in [1/2, 1], \\ \frac{2}{C}(\delta_0 + B^2C^2)^{1/2} \frac{(1+4L^{1/2}BC^{3/2})}{(1-2\alpha)^{1/2}\varphi_{3\alpha/2-1/2}(n)}, & \text{if } \alpha \in [1/3, 1/2]. \end{cases}$$

- **Discussion**
  - New result
  - Cases $\alpha = 1$ and $\alpha = 1/3$
  - Change of behavior at $\alpha = 2/3$ and $1/2$
  - Conjecture: optimal rates (for stochastic gradient descent)

Non-strongly convex functions
Averaging

- **Assumptions**: f convex, f_n' L-Lipschitz
- **Bound on** $\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]$, if $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [1/2, 1]$:

$$\frac{1}{C}\left(\delta_0 + \frac{\sigma^2}{L^2}\right)\frac{\exp\left(2L^2C^2\varphi_{1-2\alpha}(n)\right)}{n^{1-\alpha}}\left[1+(2LC)^{1+\frac{1}{\alpha}}\right]+\frac{\sigma^2 C}{2n}\varphi_{1-\alpha}(n)$$

- **Discussion**
  - Probably not new
  - Cases $\alpha = 1$ and $\alpha = 1/2$
  - Conjecture: optimal rates
  - Relationship to minimax rates

Non-strongly convex functions
Averaging - bounded gradients

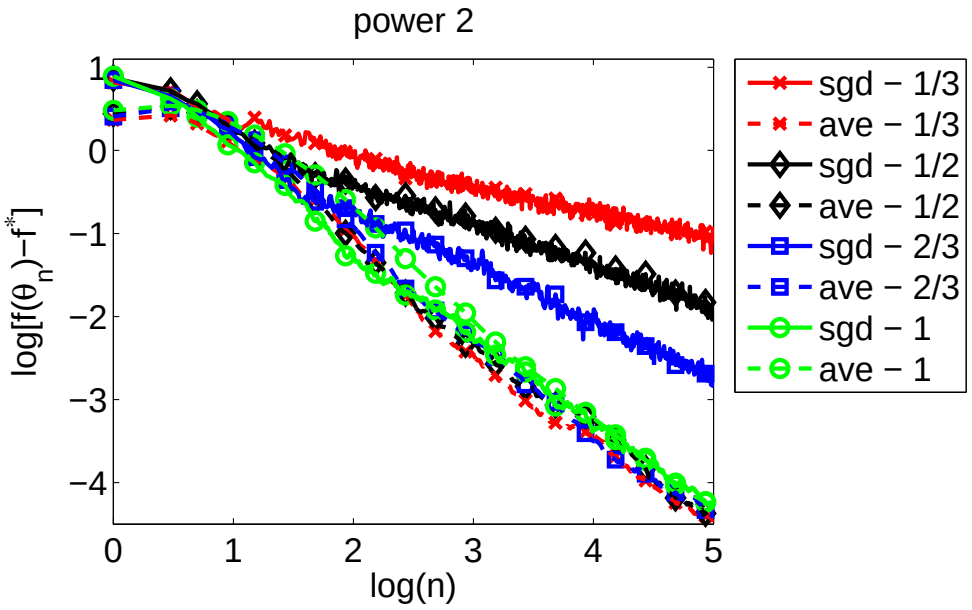
- **Assumptions**: f convex, f_n B-Lipschitz
- **Bound on** $\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]$, if $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [0, 1]$:

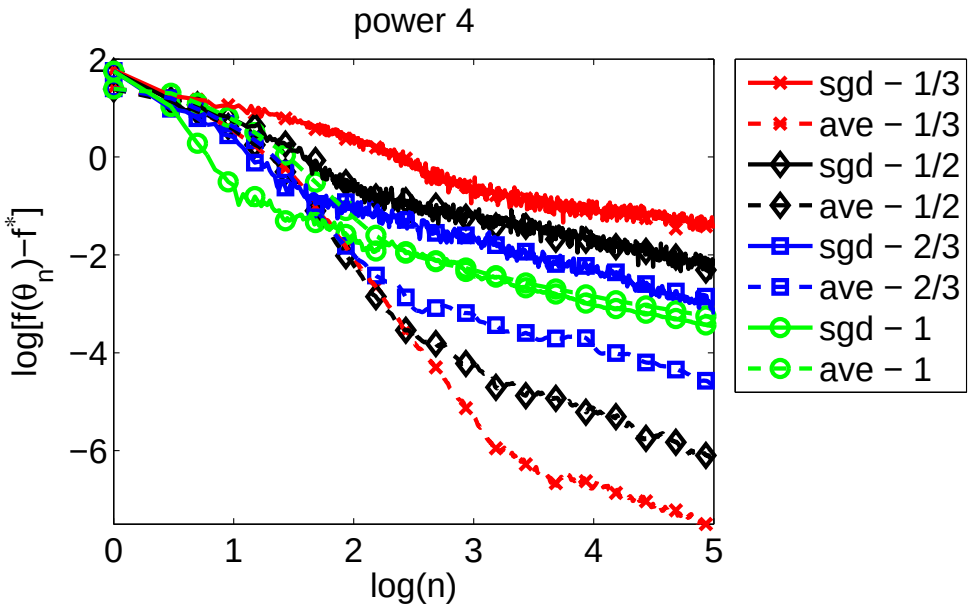
$$\frac{n^{\alpha-1}}{2C}(\delta_0 + C^2B^2\varphi_{1-2\alpha}(n)) + \frac{B^2}{2n}\varphi_{1-\alpha}(n).$$

- **Discussion**
  - Not a new result (Hazan et al., 2007; Shalev-Shwartz and Srebro, 2008; Shalev-Shwartz et al., 2007, 2009; Xiao, 2010; Duchi and Singer, 2009; Nemirovski et al., 2009)
  - Cases $\alpha = 1$ and $\alpha = 1/2$
  - Conjecture: optimal rates
  - Relationship to minimax rates

Robustness to lack of strong convexity

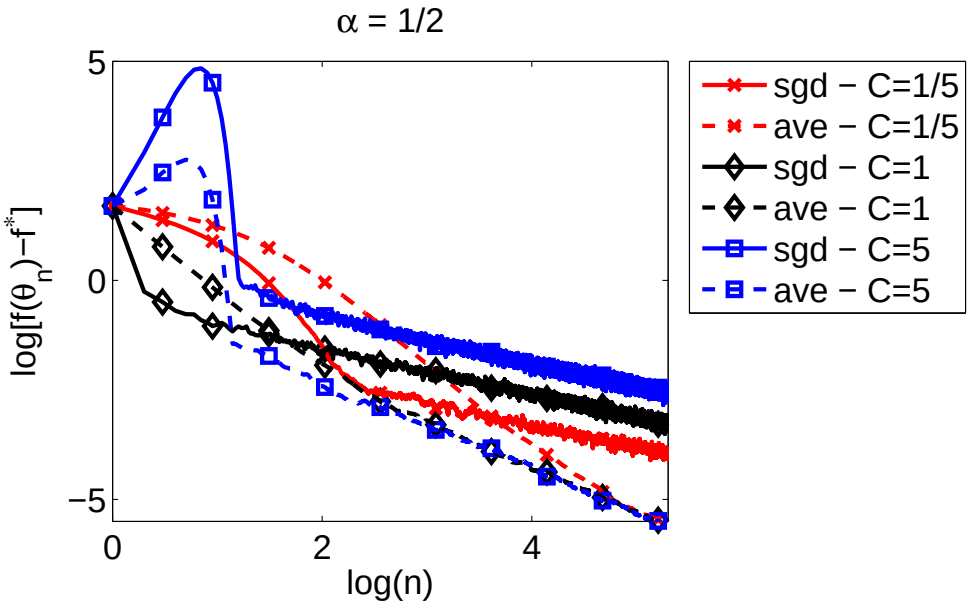
- Left: $f(\theta) = |\theta|^2$ between -1 and 1
- Right: $f(\theta) = |\theta|^4$ between -1 and 1
- affine outside of $[-1, 1]$, continuously differentiable.

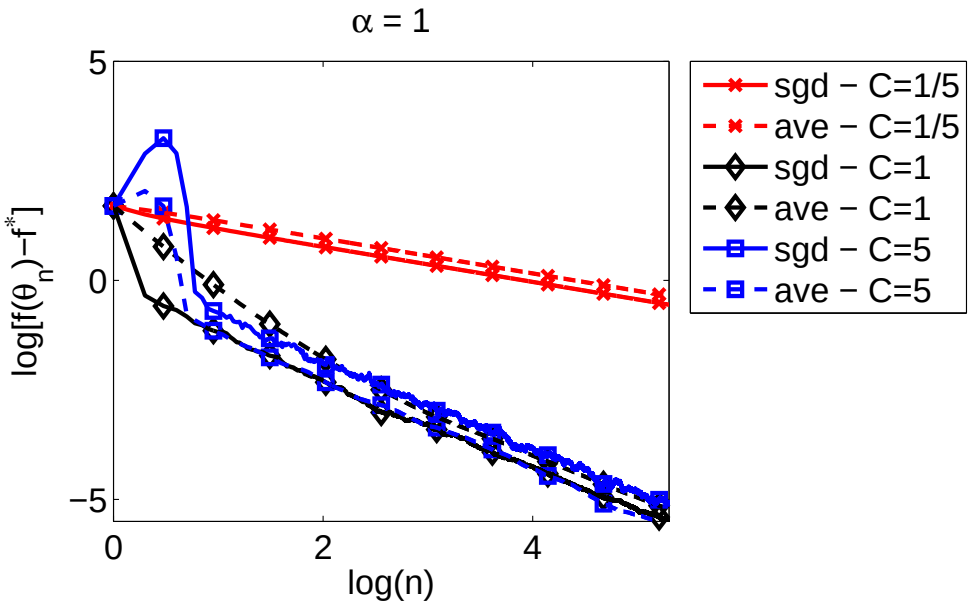




Robustness to wrong constants for $\gamma_n = Cn^{-\alpha}$

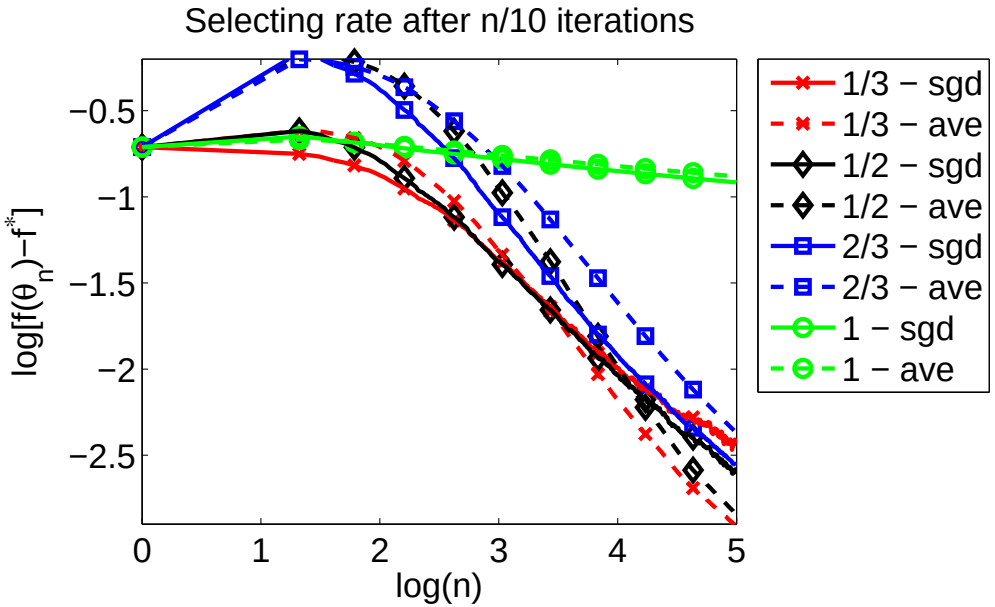
- $f(\theta) = \frac{1}{2}|\theta|^2$ with i.i.d. Gaussian noise
- Left: $\alpha = 1/2$
- Right: $\alpha = 1$

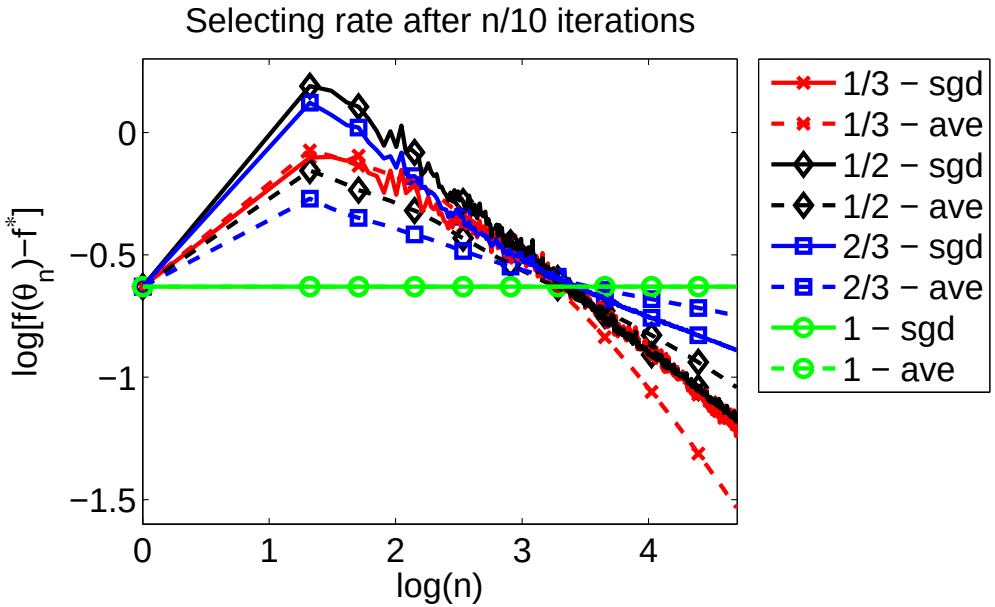




Comparison on non strongly convex logistic regression problems

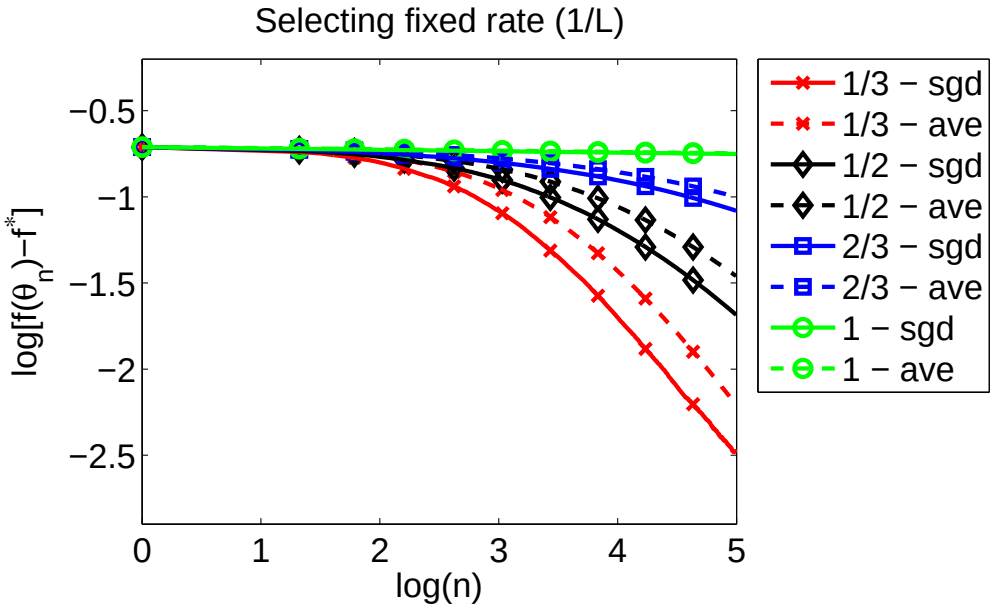
- Left: synthetic example
- Right: "alpha" dataset
- Learning constant C learned from $n/10$ iterations

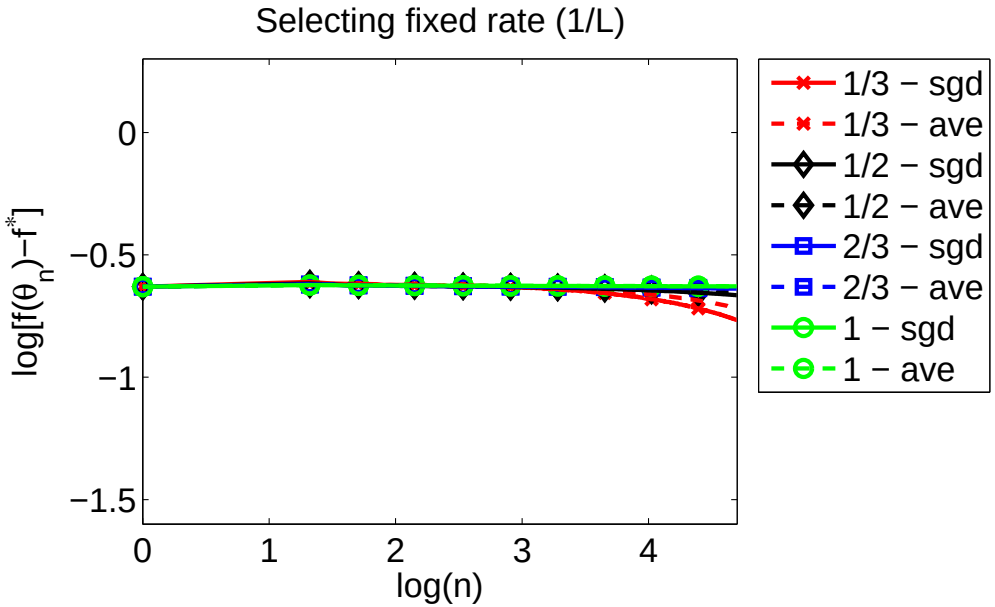




Comparison on non strongly convex logistic regression problems

- Left: synthetic example
- Right: "alpha" dataset
- Learning constant $C = 1/L$ (suggested from bounds)





Summary of results

- **Two methods**: stochastic gradient descent with and without averaging
- **Different assumptions**: μ, L, B
- **Learning rate sequence**: $\gamma_n = Cn^{-\alpha}$
- **Convergence rate of** $f(\theta_n) - f(\theta^*)$ **or** $f(\bar{\theta}_n) - f(\theta^*)$
  - $O(n^{-\beta})$, where β depends on α

α	SGD μ, L	SGD μ, B^*	SGD L	SGD L, B	Aver. μ, L	Aver. L	Aver. B
$(0 \quad , 1/3)$	α	α	$\times$	$\times$	2α	$\times$	α
$(1/3 \ , 1/2)$	α	α	$\times$	$(3\alpha - 1)/2$	2α	$\times$	α
$(1/2 \ , 2/3)$	α	α	$\alpha/2$	$\alpha/2$	1	$1 - \alpha$	$1 - \alpha$
$(2/3 \ , \quad 1)$	α	α	$1 - \alpha$	$1 - \alpha$	1	$1 - \alpha$	$1 - \alpha$

Conclusions

Stochastic approximation for machine learning

- **Mixing convex optimization and statistics**
  - Non-asymptotic analysis through moment computations
  - Averaging with longer steps is (more) robust and adaptive
  - Bounded gradient assumption leads to better rates
- **Future/current work**
  - High-probability through all moments $\mathbb{E}\|\theta_n - \theta^*\|^{2d}$
  - Analysis for logistic regression using self-concordance (Bach, 2010)
  - Including a non-differentiable term (Xiao, 2010; Lan, 2010; Duchi and Singer, 2009)
  - Non-random errors (Schmidt, Le Roux, and Bach, 2011)
  - Line search for stochastic gradient

Conclusions

Machine learning and convex optimization

- **Statistics with or without optimization?**
  - Significance of mixing algorithms with analysis
  - Benefits of mixing algorithms with analysis
- **Open problems**
  - Non-parametric stochastic approximation
  - Going beyond a single pass over the data
  - Characterization of implicit regularization of online methods
  - Further links between convex optimization and online learning/bandits

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization, 2010. Tech. report, Arxiv 1009.0571.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. ISSN 1935-7524.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Technical Report 00613125, HAL, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS), 20*, 2008.

L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

S. Boucheron and P. Massart. A high-dimensional wilks phenomenon. *Probability theory and related fields*, 150(3-4):405–433, 2011.

S. Boucheron, O. Bousquet, G. Lugosi, et al. Theory of classification: A survey of some recent advances. *ESAIM Probability and statistics*, 9:323–375, 2005.

M. N. Broadie, D. M. Cicek, and A. Zeevi. General bounds and finite-time improvement for stochastic approximation algorithms. Technical report, Columbia University, 2009.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9):2050–2057, 2004.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009. ISSN 1532-4435.

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

O. Yu. Kul′chitskiĭ and A. È. Mozgovoĭ. An estimate for the rate of convergence of recurrent robust identification algorithms. *Kibernet. i Vychisl. Tekhn.*, 89:36–39, 1991. ISSN 0454-9910.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, second edition, 2003.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, pages 1–33, 2010.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. S. Nemirovski and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep*, 76, 2007.

Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6): 1559–1568, 2008. ISSN 0005-1098.

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

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. Optimization with approximate gradients. Technical report, HAL, 2011.

S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proc. ICML*, 2008.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. ICML*, 2007.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Conference on Learning Theory (COLT)*, 2009.

K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. *Advances in Neural Information Processing Systems*, 22, 2008.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010. ISSN 1532-4435.