Stochastic gradient methods for machine learning

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with Eric Moulines, Nicolas Le Roux and Mark Schmidt - July 2012

Context

- **Large-scale machine learning**: **large p, large n, large k**
  - p : dimension of each observation (input)
  - k : number of tasks (dimension of outputs)
  - n : number of observations
- **Examples**: computer vision, bioinformatics
- **Ideal running-time complexity**: $O(pn + kn)$
- **Going back to simple methods**
  - Stochastic gradient methods (Robbins and Monro, 1951)
  - Mixing statistics and optimization
  - It is possible to improve on the rate $O(1/t)$?

Outline

- **Introduction**
  - Supervised machine learning and convex optimization
  - Beyond the separation of statistics and optimization
- **Stochastic approximation algorithms** (Bach and Moulines, 2011)
  - Stochastic gradient and averaging
  - Strongly convex vs. non-strongly convex
- **Going beyond stochastic gradient** (Le Roux, Schmidt, and Bach, 2012)
  - More than a single pass through the data
  - Exponential convergence rate

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Vector space $\mathcal{F}$ of prediction functions $\theta : \mathcal{X} \to \mathcal{Y}$
- **(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: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta(x_i))$ training cost
- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta(x))$ testing cost
- **Two fundamental questions**: (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$

Statistical analysis of empirical risk minimization

- **Error decomposition**: with $\mathcal{C} = \{\theta \in \mathcal{F}, \Omega(\theta) \leqslant D^2\}$

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

generalisation error = estimation error + approximation error

- Deviations inequalities to bound estimation error, e.g., through

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \mathcal{C}} f(\theta) &= \big[f(\hat{\theta}) - \hat{f}(\hat{\theta})\big] + \big[\hat{f}(\hat{\theta}) - \hat{f}(\theta)\big] + \big[\hat{f}(\theta) - \min_{\theta \in \mathcal{C}} f(\theta)\big] \\ &\leqslant 2 \sup_{\theta \in \mathcal{C}} |\hat{f}(\theta) - f(\theta)| \end{aligned}$$

  - See Boucheron et al. (2005); Sridharan et al. (2008); Boucheron and Massart (2011)
  - $O(1/n)$ for strongly convex functions, $O(1/\sqrt{n})$ otherwise

Iterative methods for minimizing smooth functions

- **Assumption**: f convex and smooth on $\mathcal{F}$ (Hilbert space or $\mathbb{R}^p$)

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

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

Iterative methods for minimizing smooth functions

- **Assumption**: f convex and smooth on $\mathcal{F}$ (Hilbert space or $\mathbb{R}^p$)

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

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

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

$\Rightarrow$ **Stochastic approximation**

Outline

- **Introduction**
  - Supervised machine learning and convex optimization
  - Beyond the separation of statistics and optimization
- **Stochastic approximation algorithms** (Bach and Moulines, 2011)
  - Stochastic gradient and averaging
  - Strongly convex vs. non-strongly convex
- **Going beyond stochastic gradient** (Le Roux, Schmidt, and Bach, 2012)
  - More than a single pass through the data
  - Exponential convergence rate

Stochastic approximation

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

- **Stochastic approximation**
  - Observation of $f'_n(\theta_n) = f'(\theta_n) + \varepsilon_n$
  - ε_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 gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathcal{H}$

- **Stochastic approximation**
  - Observation of $f'_n(\theta_n) = f'(\theta_n) + \varepsilon_n$
  - ε_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 $y_n \in \{-1, 1\}$.

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

- **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

- **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 $\mu > 0$:

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

Summary of new results (Bach and Moulines, 2011)

- 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
  - Forgetting of initial conditions
  - Robustness to the choice of C

- **Proof technique**
  - Derive deterministic recursion for $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$

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

  - Mimic SA proof techniques in a non-asymptotic way

Summary of new results (Bach and Moulines, 2011)

- 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 (Bach and Moulines, 2011)

- 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

Conclusions / Extensions
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 - open problems**
  - 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)
  - Non-random errors (Schmidt, Le Roux, and Bach, 2011)
  - Line search for stochastic gradient
  - Non-parametric stochastic approximation
  - Going beyond a single pass through the data

Outline

- **Introduction**
  - Supervised machine learning and convex optimization
  - Beyond the separation of statistics and optimization
- **Stochastic approximation algorithms** (Bach and Moulines, 2011)
  - Stochastic gradient and averaging
  - Strongly convex vs. non-strongly convex
- **Going beyond stochastic gradient** (Le Roux, Schmidt, and Bach, 2012)
  - More than a single pass through the data
  - Exponential convergence rate

Going beyond a single pass over the data

- **Stochastic approximation**
  - Assumes infinite data stream
  - Observations are used only once
  - Directly minimizes testing cost $\mathbb{E}_z \ell(\theta, z)$

Going beyond a single pass over the data

- **Stochastic approximation**
  - Assumes infinite data stream
  - Observations are used only once
  - Directly minimizes testing cost $\mathbb{E}_z \ell(\theta, z)$
- **Machine learning practice**
  - Finite data set $(z_1, \dots, z_n)$
  - Multiple passes
  - Minimizes training cost $\frac{1}{n}\sum_{i=1}^n \ell(\theta, z_i)$
  - Need to regularize (e.g., by the ℓ_2-norm) to avoid overfitting

Accelerating stochastic gradient - Related work

- **Momentum, gradient/iterate averaging, stochastic version of accelerated batch gradient methods**
  - Polyak and Juditsky (1992); Tseng (1998); Sunehag et al. (2009); Ghadimi and Lan (2010); Xiao (2010)
  - Can improve constants, but still have sublinear $O(1/t)$ rate
- **Constant step-size stochastic gradient (SG), accelerated SG**
  - Kesten (1958); Delyon and Juditsky (1993); Solodov (1998); Nedic and Bertsekas (2000)
  - Linear convergence, but only up to a fixed tolerance.
- **Hybrid Methods, Incremental Average Gradient**
  - Bertsekas (1997); Blatt et al. (2008)
  - Linear rate, but iterations make full passes through the data.

Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- Assume finite dataset: $\hat{f}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ and strong convexity of $\hat{f}$

- Batch gradient descent: $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n}\sum_{i=1}^{n} f_i'(\theta_{t-1})$
  - Linear (e.g., exponential) convergence rate
  - Iteration complexity is linear in n

- Stochastic gradient descent: $\theta_t = \theta_{t-1} - \gamma_t f_{i(t)}'(\theta_{t-1})$
  - $i(t)$ random element of $\{1, \dots, n\}$: sampling with replacement
  - Convergence rate in $O(1/t)$
  - Iteration complexity is independent of n

- **Best of both worlds**: linear rate with $O(1)$ iteration cost

Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient** (SAG) iteration
  - Keep in memory the gradients of all functions f_i, $i = 1, \dots, n$
  - Random selection $i(t) \in \{1, \dots, n\}$ with replacement
  - Iteration: $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^{n} y_i^t$ with $y_i^t = \begin{cases} f_i'(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$
- Stochastic version of incremental average gradient (Blatt et al., 2008)
- Extra memory requirement: same size as original data

Stochastic average gradient
Convergence analysis - I

- Assume that each f_i is L-smooth and $\frac{1}{n}\sum_{i=1}^n f_i$ is μ-strongly convex

- **Constant step size** $\gamma_t = \dfrac{1}{2nL}$:

$$\mathbb{E}\big[\|\theta_t - \theta^*\|^2\big] \leqslant \Big(1 - \frac{\mu}{8Ln}\Big)^t \Big[3\|\theta_0 - \theta^*\|^2 + \frac{9\sigma^2}{4L^2}\Big]$$

  - Linear rate with iteration cost independent of n ...
  - ... but, same behavior as batch gradient and IAG (cyclic version)

- **Proof technique**
  - Designing a quadratic Lyapunov function for a n-th order non-linear stochastic dynamical system

Stochastic average gradient
Convergence analysis - II

- Assume that each f_i is L-smooth and $\frac{1}{n}\sum_{i=1}^n f_i$ is μ-strongly convex
- **Constant step size** $\gamma_t = \dfrac{1}{2n\mu}$, if $\dfrac{\mu}{L} \geqslant \dfrac{8}{n}$

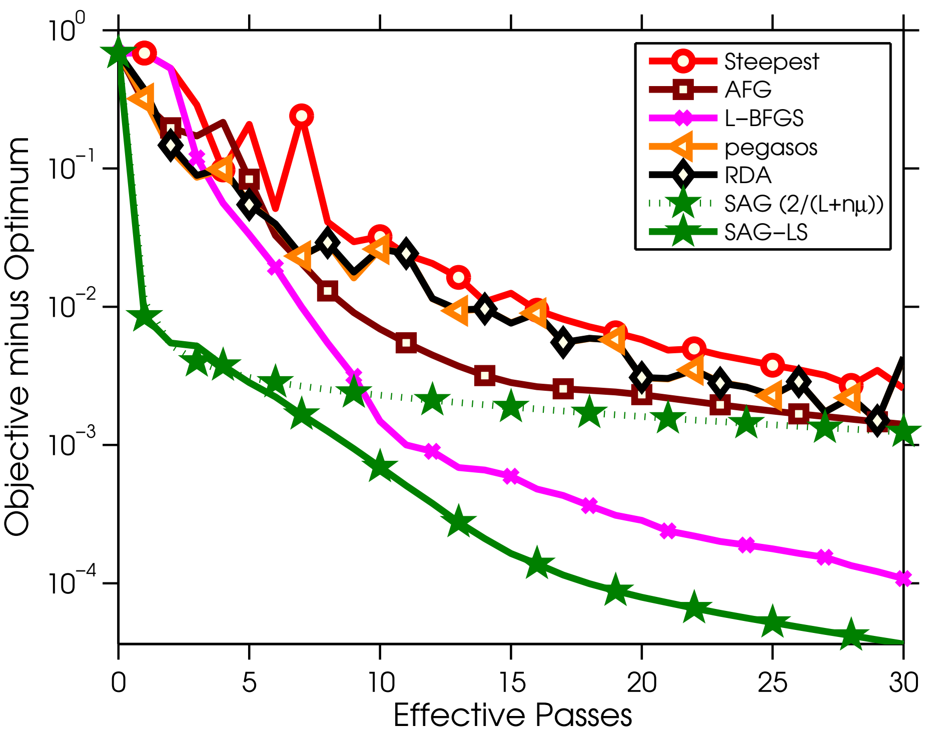
$$\mathbb{E}\big[\hat{f}(\theta_t) - \hat{f}(\theta^*)\big] \leqslant C\Big(1 - \frac{1}{8n}\Big)^t$$

with $C = \left[\frac{16L}{3n}\|\theta_0 - \theta^*\|^2 + \frac{4\sigma^2}{3n\mu}\Big(8\log\Big(1 + \frac{\mu n}{4L}\Big) + 1\Big)\right]$

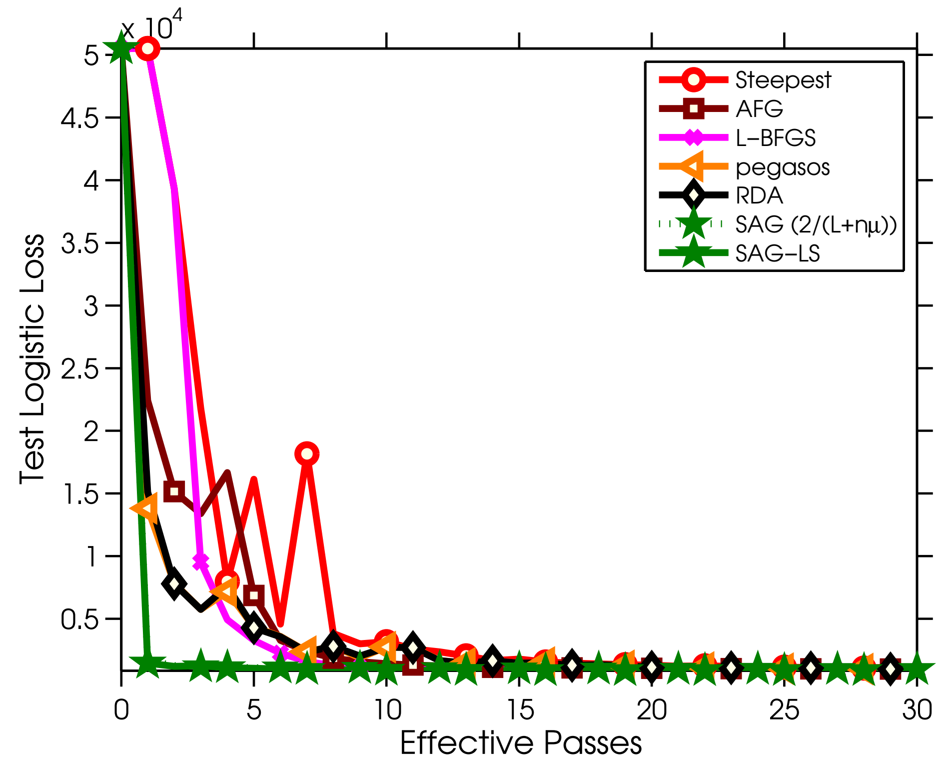
- Linear rate with iteration cost independent of n
- Linear convergence rate "independent" of the condition number
- After each pass through the data, constant error reduction

Stochastic average gradient
Simulation experiments

- `protein` dataset ($n = 145751$, $p = 74$)
- Dataset split in two (training/testing)



Training cost

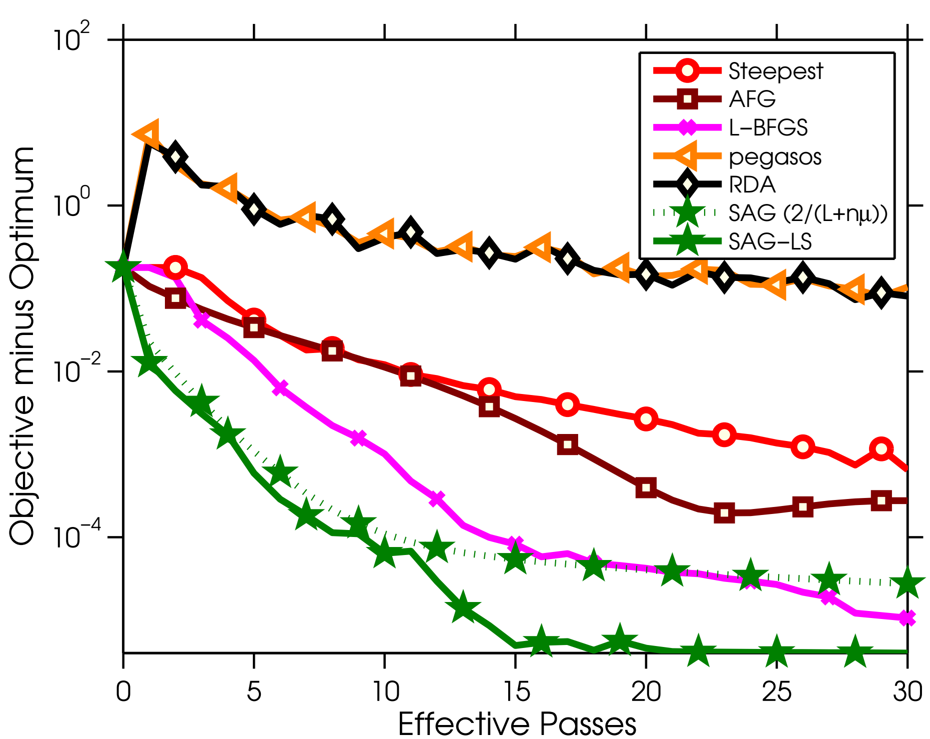


Testing cost

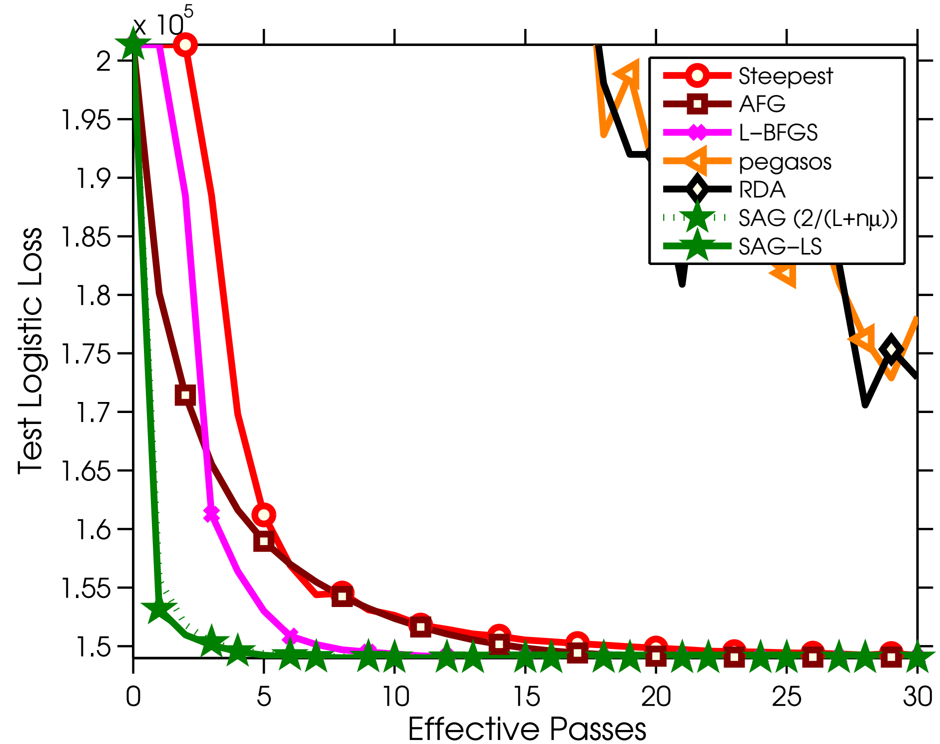
Stochastic average gradient
Simulation experiments

- `cover type` dataset ($n = 581012$, $p = 54$)

- Dataset split in two (training/testing)



Training cost



Testing cost

Conclusions / Extensions
Stochastic average gradient

- **Going beyond a single pass through the data**
  - Keep memory of all gradients for finite training sets
  - Linear convergence rate with $O(1)$ iteration complexity
  - Randomization leads to easier analysis and faster rates

- **Future/current work - open problems**
  - Including a non-differentiable term
  - Line search
  - Using second-order information or non-uniform sampling
  - Going beyond finite training sets

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 and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning, 2011.

D. P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

D. Blatt, A.O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. 18(1):29–51, 2008.

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.

B. Delyon and A. Juditsky. Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3: 868–881, 1993.

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.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization. *Optimization Online*, July, 2010.

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

H. Kesten. Accelerated stochastic approximation. *Ann. Math. Stat.*, 29(1):41–59, 1958.

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.

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. Technical Report -, HAL, 2012.

A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. *Stochastic Optimization: Algorithms and Applications*, pages 263–304, 2000.

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 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.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851.

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.

M.V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

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

P. Sunehag, J. Trumpf, SVN Vishwanathan, and N. Schraudolph. Variable metric stochastic approximation theory. *International Conference on Artificial Intelligence and Statistics*, 2009.

P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

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