Online learning and Stochastic Approximation...

Eric Moulines

Institut Mines-Télécom / Telecom ParisTech / Laboratoire Traitement et Communication de l'Information

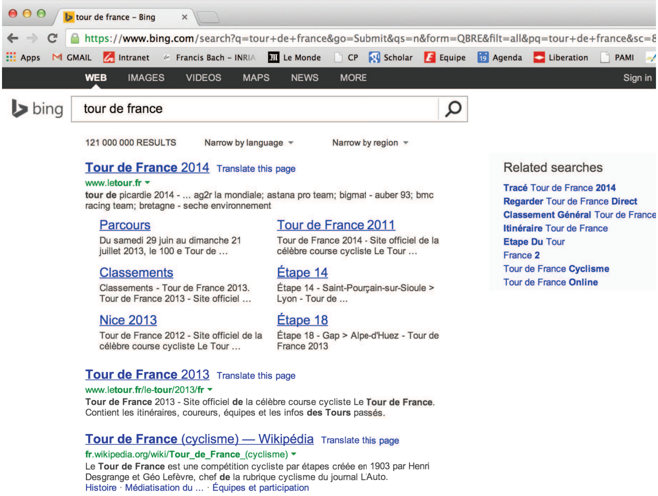
Joint work with: F. Bach, G. Fort

“Big data” revolution?

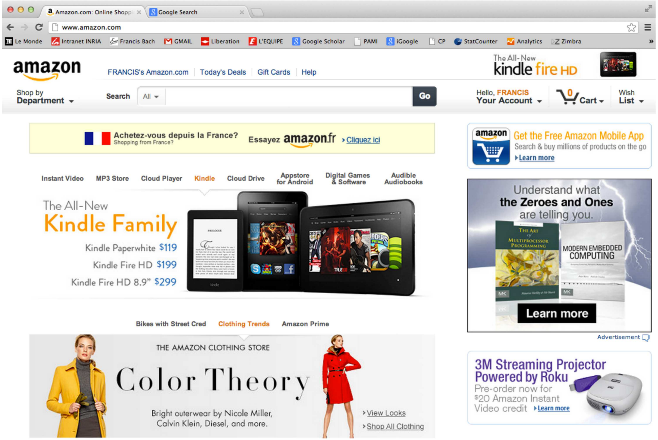
or... Why and how dull statistics become sexy ?

- Data everywhere: size does not (always) matter
- Science and industry
- Size and variety
- Learning from examples
    - n observations in dimension p

Search engines - Advertising

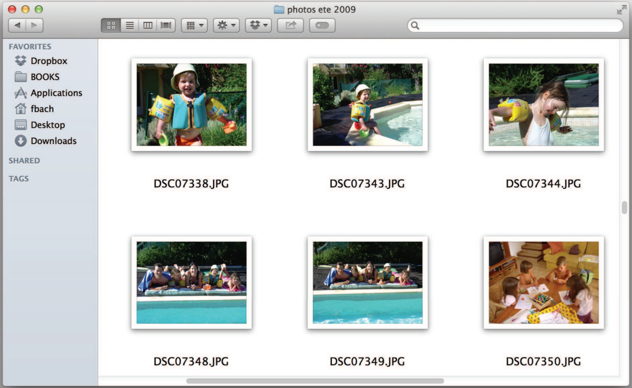


Marketing - Personalized recommendation

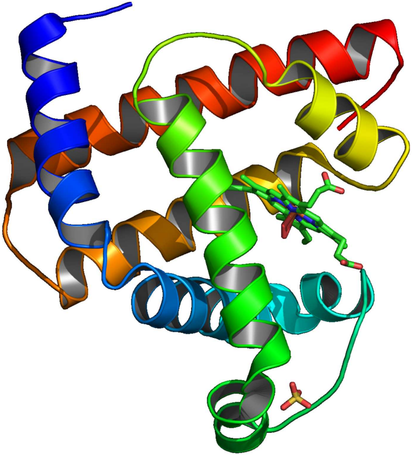


Visual object recognition

Personal photos



Bioinformatics



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

Machine learning for “big data”

- 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

  Going back to simple methods

Context

- 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

Machine learning for "big data"

- 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 !
  - Mixing statistics and optimization

Scaling to large problems with convex optimization

- 1950's: computers not powerful enough

IBM "1620", 1959
CPU frequency: 50 kHz
Price > 100000 dollars

- 2010's: Massive data !

Scaling to large problems with convex optimization

- 1950's: computers not powerful enough

IBM "1620", 1959
CPU frequency: 50 kHz
Price > 100000 dollars

- 2010's: Massive data !
- One pass through the data (Robbins et Monro, 1956)
  - Algorithm: $\boxed{\theta_n = \theta_{n-1} - \gamma_n \nabla \ell(Y_n, \langle \theta_{n-1}, \Phi(x_n) \rangle) \Phi(x_n)}$

Supervised machine learning

- Data: n observations $(X_i, Y_i) \in \mathsf{X} \times \mathsf{Y}$, $i = 1, \ldots, n$, i.i.d.
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^p$
- infinite dimensional can be dealt with as well (and implementable versions using the kernel trick are available) but this typically requires some additional care.
- (regularized) empirical risk minimization: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^p} \quad \frac{1}{N} \sum_{i=1}^{N} \ell_\theta(Y_i, X_i) \quad + \quad \mu g(\theta)$$

- data fitting term + regularizer

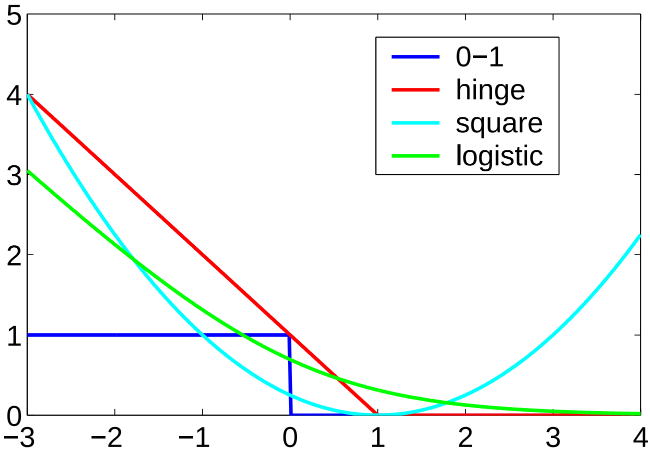
Losses for Regression

- quadratic loss: $(y \in \mathbb{R})$ $\ell_\theta(y, x) = \frac{1}{2}(y - \langle \Phi(x), \theta \rangle)^2$ where $\Phi(x)$ is a set of features.
- robust regression: $(y \in \mathbb{R})$ $\ell_\theta(y, x) = \rho(y - \langle \Phi(x), \theta \rangle)$ where ρ is a Huberized loss $\rho(t) = \log \cosh t$.
- generalized linear models $\ell_\theta(y, x) = -\langle \theta, \Phi(y, x) \rangle + Z(\theta)$, where $Z(\theta) = \int h(y) \exp(\langle \theta, \Phi(y, x) \rangle) \mathrm{d}y$. (includes multinomial regression and conditional random fields)

Losses for Classification

- "True" 0-1 loss: $\ell_\theta(y,x) = \mathbb{1}_{\{y \operatorname{sign}(\langle \theta, \Phi(x) \rangle) < 0\}}$ usually intractable
- Convexification of the $0-1$ loss are often easier to deal with and are most often used in practice...
- Hinge loss $\ell_\theta(y,x) = \max(0, 1 - y\langle \theta, \Phi(x) \rangle)$. With the penalty $g(\theta) = \|\theta\|^2$, the hinge loss is used for maximum margin classification, most notably for support vector machines
- Logistic loss $\ell_\theta(y,x) = \log(1 + \exp(-y\langle \theta, \Phi(x) \rangle))$. Taking $g(\theta) = \|\theta\|^2$ yields to ridge logistic regression.

Usual losses for classifications



Wrap-up

- Convex optimization forms the backbone of many algorithms for statistical learning and estimation.
- Given that many statistical estimation problems are large-scale in nature - problem dimension and/or sample size are large-
- It is essential to make efficient use of computational resources.
- Stochastic optimization algorithms are an attractive class of methods, known to yield moderately accurate solutions in a relatively short time

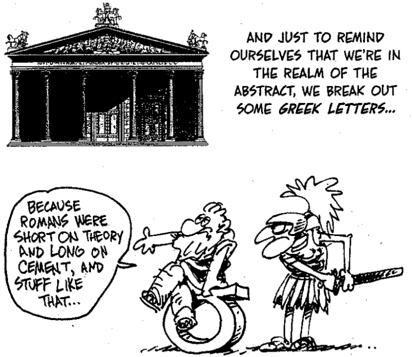
Subgradient

- The idea of derivative allows us to approximate functions by linear functions. When we minimize functions, one-sided approximation is sufficient.
- In place of the gradient we may therefore consider the subgradient, the element of Θ satisfying, for all $\theta, \vartheta \in \Theta$,

$$f(\theta) + \langle \phi, \vartheta - \theta \rangle \leq f(\vartheta)$$

- The set of subgradients (called the subdifferential) is denoted $\partial f(\theta)$.

Ghosh....


AND JUST TO REMIND OURSELVES THAT WE'RE IN THE REALM OF THE ABSTRACT, WE BREAK OUT SOME GREEK LETTERS...
BECAUSE ROMANS WERE SHORT ON THEORY AND LONG ON CEMENT, AND STUFF LIKE THAT...

Properties of the subgradient

- Terminology The domain of a function f is the set $\{\theta \in \Theta, f(\theta) < \infty\}$. The function f is convex if it is convex on its domain.

Properties of the subgradient

- Terminology The domain of a function f is the set $\{\theta \in \Theta, f(\theta) < \infty\}$. The function f is convex if it is convex on its domain.
- Properties:
  - For any proper function f, the point θ_* is a (global) minimizer of f if and only if $0 \in \partial f(\theta_*)$.
  - If $\theta \in \mathrm{int}(\mathrm{Dom}(f))$ then $\partial f(\theta) \neq \emptyset$.
  - f is Gâteaux differentiable at θ exactly when f has a unique subgradient.

The subgradient descent algorithm

- Denote by θ_* be an optimal solution of the problem $\min_{\theta \in \Theta} f(\theta)$.

$$\|\theta_{n+1} - \theta_*\|^2 \leq \|\theta_n - \theta_*\|^2 - 2\gamma_{n+1}\langle \theta_n - \theta_*, \phi_n \rangle + \gamma_{n+1}^2 \|\phi_n\|^2$$

- For any $\phi \in \partial f(\theta)$, we have

$$f(\vartheta) \geq f(\theta) + \langle \phi, \vartheta - \theta \rangle \ .$$

which implies $(\vartheta \leftarrow \theta_*, \theta \leftarrow \theta_n)$

$$0 \leq f(\theta_n) - f(\theta_*) \leq \langle \phi_n, \theta_n - \theta_* \rangle$$

The subgradient descent

- Combining the two inequalities, we obtain

$$(2\gamma_{n+1})\{f(\theta_n) - f(\theta_*)\} \leq \|\theta_n - \theta_*\|^2 - \|\theta_{n+1} - \theta_*\|^2 + \gamma_{n+1}^2 \|\phi_n\|^2 .$$

- Note that the subgradient descent is not a monotone algorithm...
- Consider the weighted averaged estimator

$$\bar{\theta}_n^{\gamma} = \Gamma_n^{-1} \sum_{k=0}^{n-1} \gamma_k \theta_k \;, \quad \Gamma_n = \sum_{k=1}^{n} \gamma_k \;.$$

- Since f is convex, $f(\bar{\theta}_n^{\gamma}) \leq \Gamma_n^{-1} \sum_{k=1}^n \gamma_k f(\theta_k)$. Assuming that $\|\phi_n\| \leq R$ (the subgradients are uniformly bounded)

$$0 \leq f(\bar{\theta}_n^{\gamma}) - f(\theta_*) \leq (2\Gamma_n)^{-1} \|\theta_0 - \theta_*\|^2 + R^2 (2\Gamma_n)^{-1} \sum_{k=0}^{n-1} \gamma_{k+1}^2 \;.$$

Fixed horizon algorithm

- For a fixed optimization horizon n, it is optimal to set $\gamma_k = \gamma$, and the previous result implies

$$0 \leq f(\bar{\theta}_n^{\gamma}) - f(\theta_*) \leq (2\gamma n)^{-1} \|\theta_0 - \theta_*\|^2 + R^2 \gamma / 2 \;.$$

- Optimizing with respect to to the stepsize γ yields to the (a bit artificial...)

$$\gamma = \frac{\|\theta_0 - \theta_*\|}{\sqrt{n}}$$

and, for this choice of γ (depending on the optimization horizon)

$$0 \leq f(\bar{\theta}_n^{\gamma}) - f(\theta_*) \leq R \|\theta_0 - \theta_*\| n^{-1/2} \;.$$

Anytime algorithm

- If we take $\gamma_n \equiv n^{-\alpha}$ with $\alpha \in [0, 1/2)$, $\Gamma_n \equiv n^{1-\alpha}$, $\Gamma_n^{-1} \sum_{k=1}^n \gamma_k^2 \equiv n^{-\alpha}$
- If we take $\gamma_n \equiv n^{-\alpha}$ with $\alpha \in [1/2, 1)$, $\Gamma_n \equiv n^{1-\alpha}$, $\Gamma_n^{-1} \sum_{k=1}^n \gamma_k^2 \equiv n^{\alpha-1}$
- The optimal choice is $\gamma_n \equiv n^{1/2}$. You might have the impression that we have to loose a $\log$ factor (who cares ?!) by using the doubling trick... divide time into periods $\left[2^k, 2^{k+1} - 1\right)$ of length 2^k and choose $\gamma_j = C 2^{-k/2}$ on each period. The anytime algorithm will then have exactly the same performance than the fixed horizon.

Projected subgradient descent

- Assume that Θ is a compact convex set and let Π be the projection on Θ. Consider the algorithm

$$\begin{aligned} \vartheta_{k+1} &= \theta_k - \gamma_{k+1}\phi_k & \phi_k \in \partial f(\theta_k) \\ \theta_{k+1} &= \Pi(\vartheta_k) \end{aligned}$$

- Since Π is a contraction,

$$\|\theta_{k+1} - \theta_*\| \leq \|\Pi(\vartheta_{k+1}) - \Pi(\theta_*)\| \leq \|\theta_k - \gamma_{k+1}\phi_k\|$$

the proof can be carried out exactly along the same lines with the additional guarantee that $\|\theta_k - \theta_*\|$ remains bounded during all the iterations (there is no guarantee of that sort otherwise).

Other averaging can be considered

- Averaging with weights γ_k is not really needed. Starting from

  $$f(\theta_n) - f(\theta_*) \leq (2\gamma_{n+1})^{-1}\{\|\theta_n - \theta_*\|^2 - \|\theta_{n+1} - \theta_*\|^2\} + (\gamma_{n+1}/2)R^2,$$

  and using again the convexity of f, we get

  $$n^{-1}\sum_{k=0}^{n-1}\{f(\theta_k) - f(\theta_*)\} \leq (n\gamma_1)^{-1}\|\theta_0 - \theta_*\| + \sum_{k=1}^{n-1}\|\theta_k - \theta_*\|(\gamma_{k+1}^{-1} - \gamma_k^{-1}) + R^2\Gamma_n/(2n)$$

- If $\|\theta_k - \theta_*\| \leq B$, taking $\gamma_k \equiv \sqrt{k}$ we obtain bounds which are similar as before.. Averaging is important, but the choice of the weights does not matter much !

What is a smooth function ?

- A function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be L-smooth if it is continuously differentiable and for all $\theta, \vartheta \in \Theta$ and if its gradient is Lipshitz

$$\|\nabla f(\theta) - \nabla f(\vartheta)\| \leq L\|\theta - \vartheta\|$$

- f is L-smooth (not necessarily convex): for all ϑ, θ,

$$f(\vartheta) \leq f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + (L/2)\|\vartheta - \theta\|^2 \ .$$

- If f is convex and differentiable, $\partial f(\theta) = \{\nabla f(\theta)\}$ and the subgradient identity shows that, for all ϑ, θ,

$$0 \leq f(\vartheta) - f(\theta) - \langle \nabla f(\theta), \vartheta - \theta \rangle$$

A characterization of L-smooth functions

Lemma

Let f be such that for all $\theta, \vartheta \in \Theta$, $0 \leq f(\vartheta) - f(\theta) - \langle \nabla f(\theta), \vartheta - \theta \rangle$ Then for any $\theta, \ \vartheta \in \mathbb{R}$,

$$f(\theta) - f(\vartheta) \leq \langle \nabla f(\theta), \theta - \vartheta \rangle - \frac{1}{2L} \| \nabla f(\theta) - \nabla f(\vartheta) \|^2$$

A characterization of L-smooth functions

Lemma

Let f be such that for all $\theta, \vartheta \in \Theta$, $0 \leq f(\vartheta) - f(\theta) - \langle \nabla f(\theta), \vartheta - \theta \rangle$ Then for any $\theta,\ \vartheta \in \mathbb{R}$,

$$f(\theta) - f(\vartheta) \leq \langle \nabla f(\theta), \theta - \vartheta \rangle - \frac{1}{2L} \| \nabla f(\theta) - \nabla f(\vartheta) \|^2$$

Let $\zeta = \vartheta - (1/L)(\nabla f(\vartheta) - \nabla f(\theta))$.

$$\begin{aligned} f(\theta) - f(\vartheta) &= f(\theta) - f(\zeta) + f(\zeta) - f(\vartheta) \\ &\leq \langle \nabla f(\theta), \theta - \zeta \rangle + \langle \nabla f(\vartheta), \zeta - \vartheta \rangle + \frac{L}{2} \| \zeta - \vartheta \|^2 \\ &= \langle \nabla f(\theta), \theta - \vartheta \rangle - \frac{1}{L} \| \nabla f(\theta) - \nabla f(\vartheta) \|^2 + \frac{1}{2L} \| \nabla f(\theta) - \nabla f(\vartheta) \|^2 \end{aligned}$$

What does it mean in practice ?

- if $f(\theta) = \mathbb{E}[\log(1 + \exp(-y\langle\theta, \Phi(x)\rangle))]$ is the logistic loss then

$$\nabla f(\theta) = \mathbb{E}\left[\frac{-Y\Phi(X)}{1 + \exp(+Y\langle\theta, \Phi(X)\rangle)}\right]$$

  and $\theta \mapsto \nabla f(\theta)$ is L-smooth provided $\mathbb{E}[\|\Phi(X)\|^2] < \infty$.
- Similar results hold for the losses functions

Gradient algorithm

- Assumption: f convex and L-smooth on $\mathbb{R}^p$
- Gradient descent: $\theta_n = \theta_{n-1} - \gamma_n \nabla f(\theta_{n-1})$
- The rationale is to make a small step in the direction that minimizes the local first-order approximation. (known as the steepest descent direction).
- Can be studied in the Majorize-Minimize (MM) framework. For a gradient Lipshitz function

$$f(\vartheta) \leq f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + (2\gamma)^{-1} \|\vartheta - \theta\|^2$$

  if $(2\gamma)^{-1} \leq L/2$. Minimizing the majorizing function yields to the gradient update.

Descent property of the gradient algorithm

- If f is convex and L-smooth, then for any $\theta,\ \vartheta \in \mathbb{R}$, one has

$$0 \leq f(\vartheta) - f(\theta) - \langle \nabla f(\theta), \vartheta - \theta \rangle \leq \frac{L}{2} \|\theta - \vartheta\|^2$$

- This gives in particular the following important inequality to evaluate the improvement in one step of gradient descent:

$$f\left(\theta - \gamma \nabla f(\theta)\right) - f(\theta) \leq -\gamma(1 - L\gamma/2) \|\nabla f(\theta)\|^2$$

- If we take $\gamma \leq 2/L$, the algorithm is monotone !

Descent property

- The characterization of L-smooth functions implies

$$f(\theta) - f(\vartheta) \leq \langle \nabla f(\theta), \theta - \vartheta \rangle - \frac{1}{2L} \|\nabla f(\theta) - \nabla f(\vartheta)\|^2$$
$$f(\vartheta) - f(\theta) \leq \langle \nabla f(\vartheta), \vartheta - \theta \rangle - \frac{1}{2L} \|\nabla f(\theta) - \nabla f(\vartheta)\|^2$$

  showing that

$$\langle \nabla f(\theta) - \nabla f(\vartheta), \theta - \vartheta \rangle \geq \frac{1}{L} \|\nabla f(\theta) - \nabla f(\vartheta)\|^2$$

- If θ_* is a stationary point, $\nabla f(\theta_*) = 0$, this inequality implies

$$\langle \nabla f(\theta), \theta - \theta_* \rangle \geq \frac{1}{L} \|\nabla f(\theta)\|^2$$

Descent property

$$
\begin{aligned}
\|\theta_{k+1} - \theta^*\|^2 &= \|\theta_k - \gamma \nabla f(\theta_k) - \theta^*\|^2 \\
&= \|\theta_k - \theta^*\|^2 - 2\gamma \langle \nabla f(\theta_k), \theta_k - \theta^* \rangle + \gamma^2 \|\nabla f(\theta_k)\|^2 \\
&\leq \|\theta_k - \theta^*\|^2 - \frac{2\gamma}{L}(1 - \gamma L/2)\|\nabla f(\theta_k)\|^2
\end{aligned}
$$

Since $\gamma \leq 2/L$, we have at the same time

$$
\begin{aligned}
& f(\theta_{k+1}) - f(\theta_*) \leq f(\theta_k) - f(\theta_*) - \gamma(1 - L\gamma/2)\|\nabla f(\theta_k)\|^2 \\
& \|\theta_{k+1} - \theta_*\| \leq \|\theta_k - \theta_*\|
\end{aligned}
$$

None of these properties were satisfied by the subgradient descent algorithm... we have specifically used the property of L-smooth functions to obtain these results.

Rate of convergence of the gradient algorithm

- Denoting $\delta_k = f(\theta_k) - f(\theta^*)$, the descent property implies :

$$\delta_{k+1} \leq \delta_k - \gamma(1 - L\gamma/2)\|\nabla f(\theta_k)\|^2$$

- The convexity and the inequality $\|\theta_k - \theta_*\| \leq \|\theta_1 - \theta_*\|$ implies

$$\begin{aligned}\delta_k &\leq \langle \nabla f(\theta_k), \theta_k - \theta^* \rangle \leq \|\theta_k - \theta^*\| \; \|\nabla f(\theta_k)\| \\ &\leq \|\theta_1 - \theta_*\| \|\nabla f(\theta_k)\|\end{aligned}$$

- Combining these two inequalities yield

$$\delta_{k+1} \leq \delta_k - \gamma(1 - L\gamma/2)\delta_k^2 / \|\theta_0 - \theta^*\|^2.$$

Rate of convergence of the gradient algorithm

Set $\omega = \gamma(1-\gamma L/2)/\|\theta_0 - \theta^*\|^2$ and recall that $\delta_k/\delta_{k+1} \geq 1$.

$$\omega \delta_k^2 + \delta_{k+1} \leq \delta_k \Leftrightarrow \omega \frac{\delta_k}{\delta_{k+1}} + \frac{1}{\delta_k} \leq \frac{1}{\delta_{k+1}} \qquad \Rightarrow \frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \omega$$
$$\Rightarrow \frac{1}{\delta_n} \geq \omega(n-1).$$

Theorem

Let f be convex and L-smooth. Then the gradient descent algorithm with $\gamma \leq 2/L$ satisfies

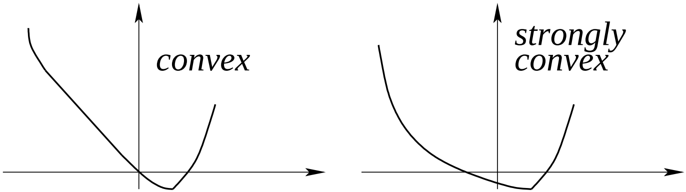
$$f(\theta_n) - f(\theta_*) \leq \frac{\|\theta_0 - \theta_*\|^2}{\gamma(1 - L\gamma/2)n} .$$

This rate may be shown to be optimal in a well-defined sense.

Strong convexity

A continuously differentiable convex function f is strongly convex if there exists a constant $\mu > 0$ such that

$$f(\vartheta) \geq f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + (1/2)\mu \|\vartheta - \theta\|^2 .$$



Strongly convex function

$$f(\vartheta) \geq f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + (1/2)\mu \|\vartheta - \theta\|^2 \; .$$

- Applying the strong convexity inequality at a stationary point θ_*,

$$f(\vartheta) \geq f(\theta_*) + (1/2)\mu \|\vartheta - \theta_*\|^2 \; .$$

- Characterization f is strongly convex if (and only if)

$$\langle \nabla f(\vartheta) - \nabla f(\theta), \vartheta - \theta \rangle \geq \mu \|\vartheta - \theta\|^2 \; .$$

Condition number of a function

If a function f is both L-smooth and gradient Lipshitz then

$$\langle \nabla f(\theta) - \nabla f(\vartheta), \theta - \vartheta \rangle \leq L \|\theta - \vartheta\|^2$$
$$\langle \nabla f(\theta) - \nabla f(\vartheta), \theta - \vartheta \rangle \geq \mu \|\theta - \vartheta\|^2 \, .$$

The value $Q_f = L/\mu$ is the condition number of the function.

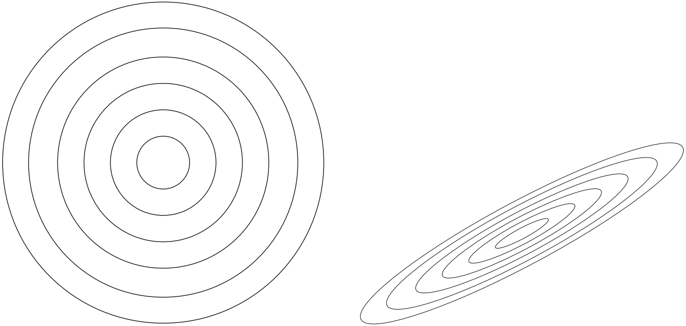


Figure: left: $\mu/L \approx 1$; right: $\mu/L \ll 1$

Twice continuously differentiable function

- A twice differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex if for all $\theta \in \mathbb{R}^p$, $\lambda_{\min}(H(\theta)) \succeq \mu$

Twice continuously differentiable function

- A twice differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex if for all $\theta \in \mathbb{R}^p$, $\lambda_{\min}(H(\theta)) \succeq \mu$
- Adding regularization by $(\mu/2)\|\theta\|^2$ introduces additional bias unless μ is small.

Strongly convex smooth functions

Recall that if f is both L-smooth and μ-strongly convex

$$\|\nabla f(\theta) - \nabla f(\vartheta)\| \leq L\|\theta - \vartheta\|$$
$$\langle \nabla f(\theta) - \nabla f(\vartheta), \theta - \vartheta \rangle \geq \mu \|\theta - \vartheta\|^2 \,.$$

Then, (plugging $\theta \leftarrow \theta_k$, $\vartheta \leftarrow \theta_*$ and using $\nabla f(\theta_*) = 0$)

$$\begin{aligned}
\|\theta_{k+1} - \theta_*\|^2 &= \|\theta_k - \gamma \nabla f(\theta_k) - \theta_*\|^2 \\
&= \|\theta_k - \theta_*\|^2 - 2\gamma \langle \nabla f(\theta_k), \theta_k - \theta_* \rangle + \gamma^2 \|\nabla f(\theta_k)\|^2 \\
&\leq (1 - 2\gamma\mu + \gamma^2 L^2) \|\theta_k - \theta_*\|^2
\end{aligned}$$

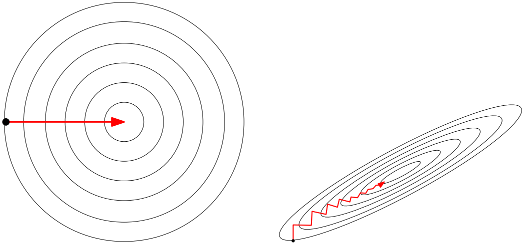
The convergence to the equilibrium is geometrically fast...

Strongly convex functions

The rate of convergence is optimized by taking $\gamma = \mu/L^2$, in which case

$$\|\theta_k - \theta_*\|^2 \leq (1-Q)^k \|\theta_1 - \theta_*\|^2$$

where $Q = \mu/L$ is the condition number of the function f.



Stochastic approximation

- Goal: Minimizing a function f defined on $\mathbb{R}^p$ given only (conditionally) unbiased estimates $\nabla f_n(\theta_n)$ of its gradients $\nabla f(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^p$
- Online learning: f is the generalization error

$$f(\theta) = \mathbb{E}[\ell_\theta(Y_1, X_1)](+g(\theta))$$

  The observations are processed as a stream (each observation is used only once and then dropped)
- Batch learning: f is the empirical risk with a complexity penalty

$$f(\theta) = N^{-1} \sum_{k=1}^{N} \ell_\theta(Y_k, X_k) + g(\theta) \; .$$

  At each iteration, we take a subsample from the training set (the observations may be therefore used several times).

Approximation of the (sub)gradients

- online case: the data are processed sequentially and

$$f_n(\theta) = m_n^{-1} \sum_{k=M_n+1}^{M_n+m_n} \ell_\theta(Y_k, X_k),$$

  where m_k is the size of the minibatches and $M_n = \sum_{k=1}^{n} m_k$. In the simplest cases, $m_n = m$ (often, $m = 1$).
- batch case: at each iteration a new minibatch of observations are sampled from the training set,

$$f_n(\theta) = m_n^{-1} \sum_{k=1}^{m_n} \ell_\theta(Y_{I_{n,k}}, X_{I_{n,k}}) \; .$$

Assumptions

- for all θ and all $\phi \in \partial f(\theta)$, $\|\phi\| \leq R$ for some known $R < \infty$. If f is differentiable θ then $\phi = \nabla f(\theta)$.
- for all n, $\Phi_{n+1} \in \partial f_n(\theta_n)$ is (conditionally) unbiased $\mathbb{E}\left[\Phi_{n+1} \,|\, \mathcal{F}_n\right] = \phi \in \partial f(\theta_n)$ and $\|\Phi_{n+1}\| \leq R$. If f_n is differentiable at θ, then $\Phi_{n+1} = \nabla f_{n+1}(\theta_n)$.

Subgradient SA (After Nemirovski and Yudin, circa 1980)

- Let Θ be a compact convex subset and Π the projection on Θ (optional). Consider the following subgradient version of the SA

$$\theta_{n+1} = \Pi(\theta_n - \gamma_{n+1} \Phi_{n+1})$$

where $\Phi_{n+1} \in \partial f_{n+1}(\theta_n)$

- Denote by θ_* be an optimal solution of the problem $\min_{\theta \in \Theta} f(\theta)$. Since Π is a contraction and $\Pi(\theta_*) = \theta_*$,

$$\|\theta_{n+1} - \theta_*\|^2 \leq \|\theta_n - \theta_*\|^2 - 2\gamma_{n+1} \langle \theta_n - \theta_*, \Phi_{n+1} \rangle + \gamma_{n+1}^2 \|\Phi_{n+1}\|^2$$

Subgradient projected SA

- Assume $\Phi_{n+1} = \phi_n + \eta_{n+1}$ where $\phi_n \in \partial f(\theta_n)$.
- For $\phi \in \partial f(\theta)$, we have

$$f(\vartheta) \geq f(\theta) + \langle \phi(\theta), \vartheta - \theta \rangle \ .$$

  which implies

$$0 \leq f(\theta_n) - f(\theta_*) \leq \langle \phi_n, \theta_n - \theta_* \rangle$$

- Combining the two inequalities, we obtain

$$0 \leq (2\gamma_{n+1})\{f(\theta_n) - f(\theta_*)\} \leq \|\theta_n - \theta_*\|^2 - \|\theta_{n+1} - \theta_*\|^2 \\ - 2\langle \eta_{n+1}, \theta_n - \theta_* \rangle + \gamma_{n+1} R^2$$

Averaging

- Consider the weighted averaged estimator (other forms of averaging are possible)

$$\bar{\theta}_n^{\gamma} = \Gamma_n^{-1} \sum_{k=0}^{n-1} \gamma_{k+1} \theta_k \;, \quad \Gamma_n = \sum_{k=1}^{n} \gamma_k \;.$$

- Since f is convex,

$$0 \leq f(\bar{\theta}_n^{\gamma}) - f(\theta_*) \leq (2\Gamma_n)^{-1} \|\theta_0 - \theta_*\|^2 \\ + 2\Gamma_n^{-1} \sum_{k=0}^{n-1} \gamma_{k+1} \langle \eta_{k+1}, \theta_k - \theta_* \rangle + \Gamma_n^{-1} \sum_{k=0}^{n-1} \gamma_{k+1}^2 \|\Phi_{k+1}\|^2 \;.$$

Moment bound

- Under the assumptions $\mathbb{E}\left[\eta_{n+1} \,|\, \mathcal{F}_n\right] = 0$ and that $\|\Phi_{n+1}\|^2 \leq R^2$ (constant mini-batches size),

$$0 \leq \mathbb{E}[f(\bar{\theta}_n^{\gamma})] - f(\theta_*) \leq (2\Gamma_n)^{-1}\mathbb{E}[\|\theta_0 - \theta_*\|^2] + \Gamma_n^{-1} R^2 \sum_{k=0}^{n-1} \gamma_{k+1}^2 \,.$$

- Assuming that $\gamma_n \sim C n^{-\alpha}$ with $\alpha \in [0,1)$ we get

$$\mathbb{E}[f(\bar{\theta}_n^{\gamma})] - f(\theta_*) \leq \begin{cases} C_\alpha n^{\alpha-1} + D_\alpha \sigma^2 n^{-\alpha} & \alpha < 1/2 \\ F_\alpha n^{\alpha-1} & \alpha > 1/2 \\ G_\alpha \log(n) n^{-1/2} & \alpha = 1/2 \end{cases}$$

Convex stochastic approximation

- Key assumption: f is L-smooth and/or μ-strong convexity
- Key algorithm: stochastic gradient descent (a.k.a. Robbins-Monro)

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

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

Key recursion

- Let θ_* be the unique global minimizer of f.
- Use $V(\theta) = \|\theta - \theta_*\|^2$ as a Lyapunov function. We get

$$\|\theta_{n+1} - \theta_*\|^2 = \|\theta_n - \theta_*\|^2 - 2\gamma_{n+1}\langle \theta_n - \theta_*, \nabla f_{n+1}(\theta_n)\rangle \\ + \gamma_{n+1}^2 \|\nabla f_{n+1}(\theta_n)\|^2 .$$

- Write

$$\nabla f_{n+1}(\theta_n) = \nabla f(\theta_n) + \eta_{n+1}$$

In the online context, $\mathbb{E}\left[\eta_{n+1} \,|\, \mathcal{F}_n\right] = 0 \quad \mathbb{P} - \text{a.s.}$, but other scenarios (we will see later that other assumptions are sometimes required).

Key recursion

- Compute the conditional expectation of the two sides of the previous equation.

$$\begin{aligned}\|\theta_{n+1}-\theta_*\|^2 \leq \|\theta_n-\theta_*\|^2 - 2\gamma_{n+1}\langle\theta_n-\theta_*, \nabla f(\theta_{n-1}) - \nabla f(\theta_*)\rangle \\ - 2\gamma_{n+1}\langle\theta_n-\theta_*, \eta_{n+1}\rangle + \gamma_{n+1}^2\|\nabla f_{n+1}(\theta_n)\|^2 \ .\end{aligned}$$

- Plug the lower bound for the scalar product using the key strongly convex inequality

$$\begin{aligned}\|\theta_{n+1}-\theta_*\|^2 \leq \|\theta_n-\theta_*\|^2 - 2\gamma_n\mu\|\theta_n-\theta_*\|^2 \\ - 2\gamma_{n+1}\langle\theta_n-\theta_*, \eta_{n+1}\rangle + \gamma_{n+1}^2\|\nabla f_{n+1}(\theta_n)\|^2\end{aligned}$$

Upper bound

- Using $\nabla f_{n+1}(\theta_n) = \nabla f(\theta_{n+1}) + \eta_{n+1}$ and $\|x+y\|^2 \leq p\|x\|^2 + q\|y\|^2$, $p^{-1} + q^{-1} = 1$,

$$\|\nabla f_{n+1}(\theta_n)\|^2 \leq p\|\nabla f(\theta_n)\|^2 + q\|\eta_{n+1}\|^2$$

- Since $\|\nabla f(\vartheta) - \nabla f(\theta)\| \leq L\|\vartheta - \theta\|$, this yields

$$\begin{aligned}\|\theta_{n+1} - \theta_*\|^2 \leq (1 - 2\gamma_{n+1}\mu + p\gamma_{n+1}^2 L^2)\|\theta_n - \theta_*\|^2 \\ - 2\gamma_{n+1}\langle \theta_n - \theta_*, \eta_{n+1}\rangle + q\gamma_{n+1}^2\|\eta_{n+1}\|^2 \ .\end{aligned}$$

Wrap-up

- Assuming that $\{\gamma_n,\ n \in \mathbb{N}\}$ is nonincreasing and that p is small enough so that, for some $\delta > 0$, $2\mu - p\gamma_1^2 L^2 \geq \kappa > 0$, we get

$$\|\theta_{n+1} - \theta_*\|^2 \leq (1 - \kappa\gamma_{n+1})\|\theta_n - \theta_*\|^2 \\ - 2\gamma_{n+1}\langle \theta_n - \theta_*, \eta_{n+1}\rangle + q\gamma_{n+1}^2\|\eta_{n+1}\|^2 .$$

- This is a non-homogeneous autoregressive sequence which can be studied explicitly and from which can be deduced a variety of results (see Nemirovski, Juditsky, Lan, Shapiro, 2009 and Bach, Moulines 2011 expanded in Bach 2013)

Moment bounds

- Assume first that $\mathbb{E}\left[\eta_{n+1} \,|\, \mathcal{F}_n\right] = 0$ and $\mathbb{E}\left[\|\eta_{n+1}\|^2 \,\middle|\, \mathcal{F}_n\right] \leq R^2$, which makes perfectly sense for online learning...
- Setting $\delta_n = \gamma_n^{-1}\mathbb{E}[\|\theta_n - \theta_*\|^2]$, we get

$$\delta_{n+1} \leq (1 - \kappa\gamma_{n+1})(\gamma_n/\gamma_{n+1})\delta_n + qR^2\gamma_{n+1}$$

- Iterating the previous inequality n times

$$\delta_n \leq (\gamma_1/\gamma_n)\prod_{k=1}^{n}(1-\kappa\gamma_k)\delta_0 + qR^2\sum_{k=1}^{n}\prod_{i=k+1}^{n}(1-\kappa\gamma_i)\gamma_k .$$

- The quadratic risk is therefore a sum of two terms:
  - a transient term, depending only on the initial condition δ_0
  - a stationary term depending only on the noise variance, accounting for the fluctuation of the estimate after the extinction of the transient.

Transient term

- The simple bound $1 + t \leqslant \exp(t)$ for any $t \in \mathbb{R}$ yield

$$\prod_{k=1}^{n} (1 - \kappa \gamma_k) \le \exp\left(-\kappa \sum_{k=1}^{n} \gamma_k \right) .$$

- To forget the initial condition, it is therefore required that $\sum_{k=1}^{n} \gamma_k = \infty$.
- The critical regime is $\gamma_k = C k^{-1}$; in such case, $\sum_{k=1}^{n} \gamma_k \sim C \log(n)$ and the rate at which the transient is forgotten is typically $\sim n^{1-\kappa C}$... the choice of C is crucial !
- If $\gamma_n \sim C n^{-\alpha}$ with $\alpha < 1$, then the forgetting is $\sim n^{\alpha} \exp(-\kappa C/(1+\alpha) n^{(1-\alpha)})$.
- The forgetting of the initial condition suggests to take a *small* α.... but of course, there is a trade-off between the transient and the stationary regime !

Stationary regime

- To study the stationary regime, we must control the sum

$$\sum_{k=1}^{n} \prod_{i=k+1}^{n} (1-\kappa\gamma_i)\gamma_k \\ = \kappa^{-1} \sum_{k=1}^{n} \left\{ \prod_{i=k+1} (1-\kappa\gamma_i) - \prod_{i=k}^{n} (1-\kappa\gamma_i) \right\} \leq \kappa^{-1}$$

- This bound yields immediately to the following explicit bound

$$\gamma_n^{-1}\delta_n \leq \gamma_n^{-1} \exp\left(-\kappa \sum_{k=1}^{n} \gamma_k\right) \delta_0 + q\kappa^{-1} \; .$$

- Provides an explicit bound of convergence and a rate.

Optimizing the rate of convergence

- The optimal rate is achieved when $\gamma_n = Cn^{-1}$, for a constant C which should be chosen sufficiently large so that the transient term vanishes (in practice this requires to know μ and L when the problem is smooth).
- Any $\gamma_n \equiv n^{-\alpha}$ with $\alpha \in [0, 1)$ yield converging sequence (there is no need to assume that $\alpha > 1/2$!) Nevertheless, the rate of convergence is no longer optimal (on the other hand, a prior knowledge of μ is not required !)

The Polyak-Ruppert idea

$$
\begin{aligned}
\theta_{n+1} &= \theta_n - \gamma_{n+1} \nabla f_n(\theta_n) \\
&= \theta_n - \gamma_{n+1} \nabla f(\theta_n) - \gamma_{n+1} \eta_{n+1} \\
&= \theta_n - \gamma_{n+1} H(\theta_*)(\theta_n - \theta_*) - \gamma_{n+1} \tilde{\eta}_{n+1}
\end{aligned}
$$

where $\eta_{n+1} = \eta_{n+1} + \nabla f(\theta_n) - H(\theta_*)(\theta_n - \theta_*)$.

The Polyak-Ruppert idea

$$
\begin{aligned}
\theta_{n+1} &= \theta_n - \gamma_{n+1} \nabla f_n(\theta_n) \\
&= \theta_n - \gamma_{n+1} \nabla f(\theta_n) - \gamma_{n+1} \eta_{n+1} \\
&= \theta_n - \gamma_{n+1} H(\theta_*)(\theta_n - \theta_*) - \gamma_{n+1} \tilde{\eta}_{n+1}
\end{aligned}
$$

where $\eta_{n+1} = \eta_{n+1} + \nabla f(\theta_n) - H(\theta_*)(\theta_n - \theta_*)$. Summing up the previous expressions yields to

$$
\bar{\theta}_n - \theta_* = \frac{H^{-1}(\theta_*)}{n+1} \sum_{k=0}^{n} \gamma_{k+1}^{-1}(\theta_k - \theta_{k+1}) - \frac{H^{-1}(\theta_*)}{n+1} \sum_{k=0}^{n} \eta_{k+1} \ ,
$$

where $\bar{\theta}_n = (n+1)^{-1} \sum_{k=0}^{n} \theta_k$

Negligibility

- Summing by parts,

$$\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\theta_{k-1}-\theta_k)=\frac{1}{n}\sum_{k=1}^{n-1}(\theta_k-\theta_*)(\gamma_{k+1}^{-1}-\gamma_k^{-1}) \\ -\frac{1}{n}(\theta_n-\theta_*)\gamma_n^{-1}+\frac{1}{n}(\theta_0-\theta_*)\gamma_1^{-1},$$

- Since $\mathbb{E}[\|\theta_k-\theta_*\|]\leq C\gamma_n^{1/2}$ (for $\gamma_n\equiv n^{-\alpha}$ and $\alpha\in(0,1)$), the dominant term is of order $n^{-1}\gamma_n^{-1/2}$.
- If $\gamma_n^{-1}=o(n)$, this term is $o(n^{-1/2})$...

Leading term

- The leading term is therefore

$$\frac{H^{-1}(\theta_*)}{n+1} \sum_{k=0}^{n} \eta_{k+1}$$

- The variance of this term is of order $O(n^{-1})$... Non asymptotic control can be of course obtained...
- Summary: Averaging always leads to $\mathbb{E}[\|\bar{\theta}_n - \theta_*\|^2] \leq Cn^{-1}$ as soon as $\lim_{n\to\infty}(n\gamma_n)^{-1} + \gamma_n = 0$
- Contrary to the non-averaged case (optimal rate but μ should be known), averaging with stepsizes $\gamma_n \equiv n^{-\alpha}$ yields optimal convergence even when μ is unknown (adaptivity).
- A more refined analysis suggests to take $\gamma_n \equiv n^{-2/3}$ and allows to obtain a nonasymptotic control (Bach and Moulines (2011) gives an explicit bound, but the bound has a suboptimal dependence on μ).

Back to the function

- Since f is gradient Lipshitz, $0 \leq f(\vartheta) - f(\theta_*) \leq (L/2)\|\vartheta - \theta_*\|^2$, the averaged estimator satisfies:

$$0 \leq \mathbb{E}[f(\bar{\theta}_n)] - f(\theta_*) \leq Cn^{-1}$$

- This rate cannot be improved... take $f(\theta) = (1/2)\|\theta\|^2$ and $\{\eta_k,\ k \in \mathbb{N}\}$ an i.i.d. sequence !

Constant step-size smooth SA

- Assuming that $\{\eta_n,\ n \in \mathbb{N}\}$ is i.i.d. the recursion

$$\theta_n = \theta_{n-1} - \gamma[\nabla f(\theta_{n-1}) + \eta_n]$$

  defines an (homogeneous) Markov chain.

- Integrating the inequality

$$\|\theta_{n+1} - \theta_*\|^2 \leq (1-\kappa\gamma)\|\theta_n - \theta_*\|^2 - 2\gamma\langle\theta_n - \theta_*, \eta_{n+1}\rangle + q\gamma^2\|\eta_{n+1}\|^2 \ .$$

  yields the Foster-Lyapunov drift condition

$$P_\gamma V(\theta) \leq \lambda V(\theta) + b\ , \quad P_\gamma V(\theta) = \mathbb{E}[V(\theta - \gamma\nabla f(\theta) + \gamma\eta_1)]\ .$$

  with the drift function $V(\theta) = \|\theta - \theta_*\|^2$.

- If in addition the distribution of η_{n+1} has a positive density with respect to the Lebesgue measure, then P is a strong-Feller Markov kernel ($x \mapsto P(x, A)$ is a continuous function).

Constant step-size smooth SA

- The Markov kernel P_γ is geometrically ergodic and converges geometrically fast to its unique stationary distribution π_γ.
- When ∇f is not linear, $\int \theta \pi_\gamma(\mathrm{d}\theta) \neq \theta_* = 0$
- Ergodic theorem
  - averaged iterates converge to $\bar{\theta}_\gamma \neq \theta_*$ at rate $O(1/n)$.
  - For any $\gamma > 0$, $\int \pi_\gamma(\mathrm{d}\theta) \nabla f(\theta) = 0$
  - moreover, $\|\theta_* - \bar{\theta}_\gamma\| = O(\gamma)$.

Least-mean-square algorithm

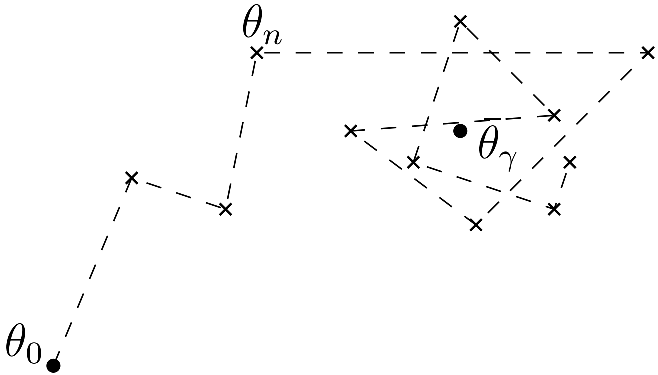
- Least-squares: $f(\theta) = (1/2)\mathbb{E}[(Y_n - \langle \Phi(X_n), \theta \rangle)^2]$ with $\theta \in \mathbb{R}^p$
- strong convexity: $\mathbb{E}\big[\Phi(X_n) \otimes \Phi(X_n)\big] = H \succcurlyeq \mu \cdot \mathrm{Id}$
- recursion $f_n(\theta) = (1/2)\big(Y_n - \langle \Phi(X_n), \theta \rangle\big)^2$

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

- $\{\theta_n,\ n \in \mathbb{N}\}$ is a (homogeneous) Markov chain a geometrically ergodic Markov chain with stationary distribution π_γ and

$$\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)\ .\bar{\theta}_\gamma = \theta_*\ .$$

Least-Mean Square

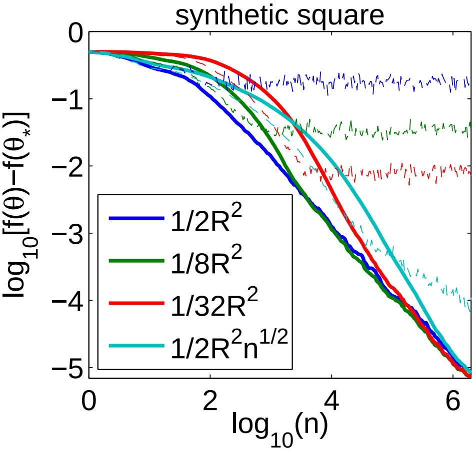


Least Mean Square

- As $n \to \infty$, the distribution of θ_n converges to π_γ which is centered around θ_*.
- By the Birkhoff theorem, $\bar{\theta}_n$ converges almost surely to θ_*.
- But of course, much more can be said... the CLT for Markov chain immediately shows that $\sqrt{n}(\bar{\theta}_n - \theta_*) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$... but non asymptotic deviation bounds can be obtained as well. item 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: $\boxed{\mathbb{E}[f(\bar{\theta}_{n-1})] - f(\theta_*) \leqslant \frac{4\sigma^2 p}{n} + \frac{4R^2\|\theta_0 - \theta_*\|^2}{n}}$
- Matches statistical lower bound

Toy Example

- Gaussian distributions - $p = 20$



Simulations - benchmarks

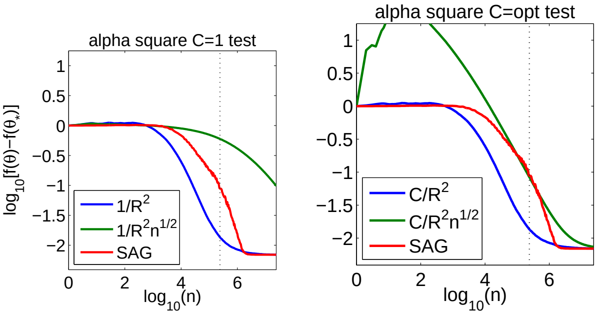


Figure: *alpha* ($p = 500$, $n = 500\ 000$)

Simulations - benchmarks

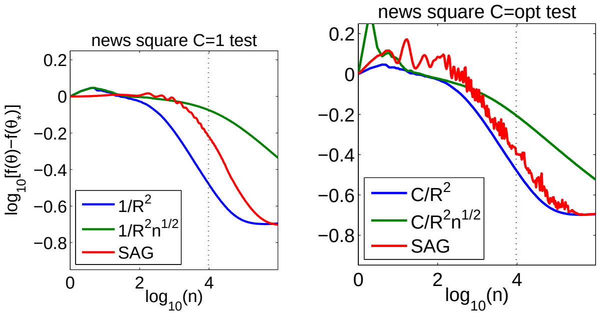


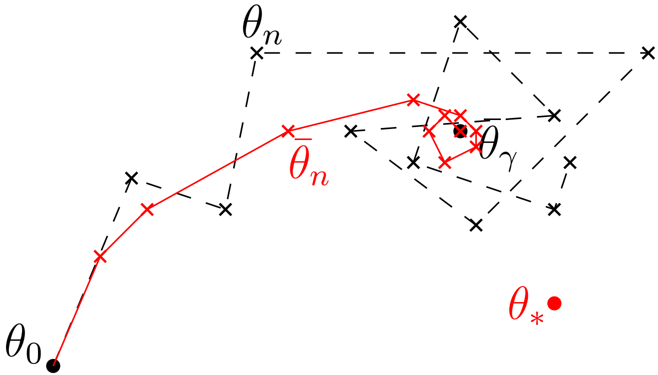
Figure: *news* ($p = 1\ 300\ 000$, $n = 20\ 000$)

Beyond Least-Mean-Squares

Recursion $\theta_n = \theta_{n-1} - \gamma \nabla f_n(\theta_{n-1})$ also defines a Markov chain (functional autoregressive)

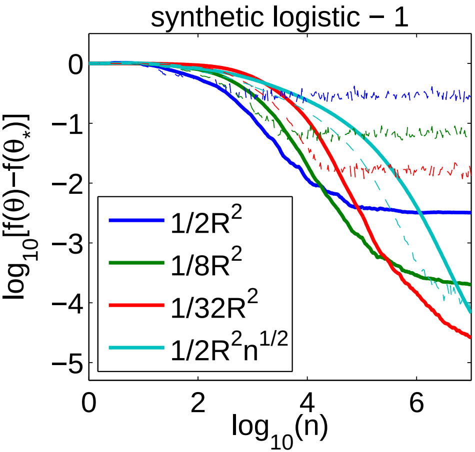
- Under appropriate conditions: Stationary distribution π_γ such that $\int \nabla f(\theta) \pi_\gamma(\mathrm{d}\theta) = 0$
- When ∇f is not linear, $\nabla f(\int \theta \pi_\gamma(\mathrm{d}\theta)) \neq \int \nabla f(\theta) \pi_\gamma(\mathrm{d}\theta) = 0$
- (θ_n) fluctuates around the wrong value $\bar{\theta}_\gamma \neq \theta_*$

Beyond Least-Mean-Squares



Toy example

- 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
  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
- 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 $f(\theta) = \mathbb{E}\big[\ell(Y_1, \langle\theta, \Phi(X_n)\rangle)\big]$ 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., **?**]
  - 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., **?**]
    - 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

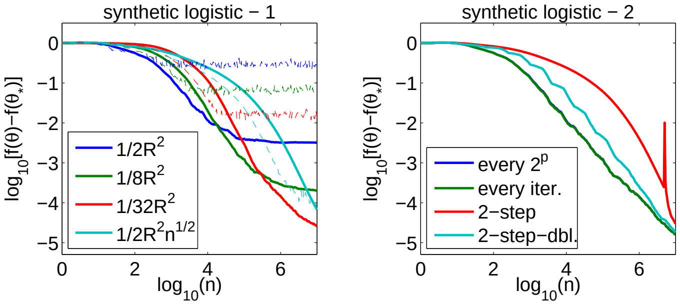
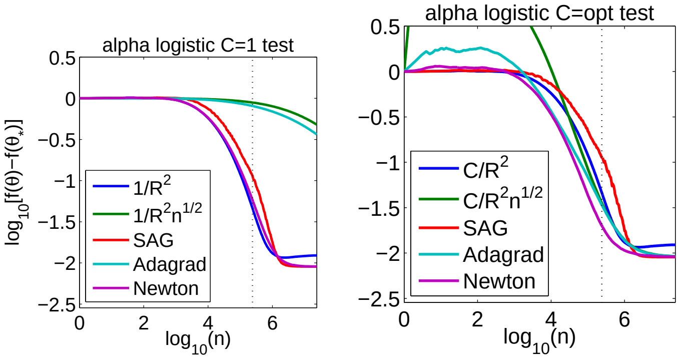


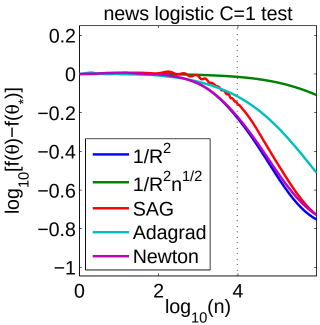
Figure: Gaussian distributions - $p = 20$

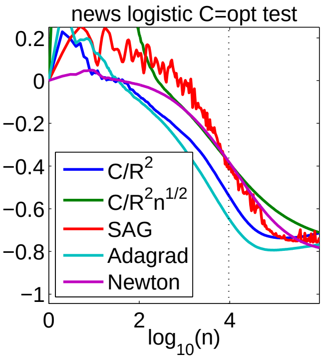
Simulations - benchmarks

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



Simulations - benchmarks





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
- Extensions and future work
  - Going beyond a single pass
  - Pre-conditioning
  - Proximal extensions for non-differentiable terms
  - kernels and non-parametric estimation [**?**]
  - line-search
  - parallelization
  - Non-convex problems

References