Bridging the gap between Stochastic Approximation and Markov chains

Aymeric DIEULEVEUT

ENS Paris, INRIA

17 november 2017

Joint work with Francis Bach and Alain Durmus.

Outline

- **Introduction to Stochastic Approximation for Machine Learning.**
- **Markov chain: a simple yet insightful point of view on constant step size Stochastic Approximation.**

Supervised Machine Learning

- Consider an input/output pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, following some unknown distribution ρ.
- $\mathcal{Y} = \mathbb{R}$ (regression) or $\{-1, 1\}$ (classification).
- We want to find a function $\theta : \mathcal{X} \to \mathbb{R}$, such that $\theta(X)$ is a good prediction for Y.
- Prediction as a linear function $\langle \theta, \Phi(X) \rangle$ of features $\Phi(X) \in \mathbb{R}^d$.
- Consider a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$: squared loss, logistic loss, 0-1 loss, etc.
- We define the risk (generalization error) as

$$\mathcal{R}(\theta) := \mathbb{E}_\rho \left[\ell(Y, \langle \theta, \Phi(X) \rangle) \right].$$

Empirical Risk minimization (I)

- Data: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, i.i.d.
  - n very large, up to 10^9
  - Computer vision: $d = 10^4$ to 10^6
- Empirical risk (or training error):

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \Phi(x_i) \rangle).$$

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

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

convex data fitting term + regularizer

Empirical Risk minimization (II)

- For example, least-squares regression:

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

- and logistic regression:

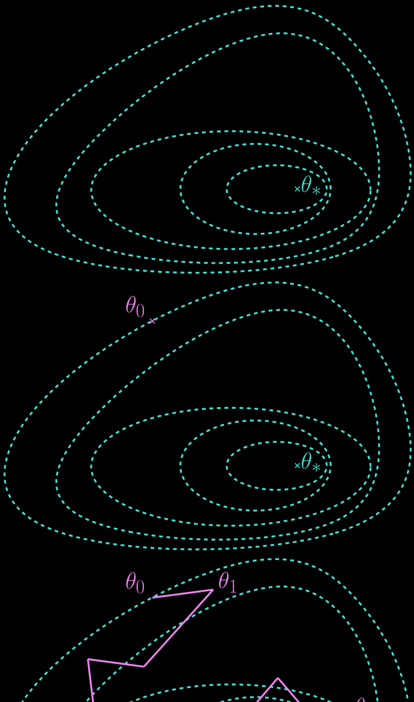
$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \log \big(1 + \exp(-y_i \langle \theta, \Phi(x_i) \rangle)\big) \quad + \quad \mu \Omega(\theta).$$

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

2 important insights for ML Bottou and Bousquet (2008):

1. No need to optimize below statistical error,
2. Testing error is more important than training error.

Stochastic Approximatic



- **Goal:**

$$\min_{\theta \in \mathbb{R}^d} f(\theta)$$

given unbiased gradient estimates f'_n

- $\theta_* := \operatorname{argmin}_{\mathbb{R}^d} f(\theta)$.

Stochastic Approximation in Machine learning

Loss for a single pair of observations, for any $k \leq n$:

$$f_k(\theta) = \ell(y_k, \langle \theta, \Phi(x_k) \rangle).$$

- Use one observation at each step !
- Complexity: $O(d)$ per iteration.
- Can be used for both true risk and empirical risk.

Stochastic Approximation in Machine learning

- For the **empirical error** $\hat{\mathcal{R}}(\theta) = \frac{1}{n}\sum_{k=1}^{n} \ell(y_k, \langle \theta, \Phi(x_k)\rangle)$.
  - At each step $k \in \mathbb{N}^*$, sample $I_k \sim \mathcal{U}\{1, \dots n\}$.
  - $\mathcal{F}_k = \sigma((x_i, y_i)_{1\le i\le n}, (I_i)_{1\le i\le k})$.
  - At step $k \in \mathbb{N}^*$, use:

$$f'_{I_k}(\theta_{k-1}) = \ell'(y_{I_k}, \langle \theta_{k-1}, \Phi(x_{I_k})\rangle)$$

$$\mathbb{E}[f'_{I_{I_k}}(\theta_{k-1})|\mathcal{F}_{k-1}] = \hat{\mathcal{R}}'(\theta_{k-1})$$

- For the **risk** $\mathcal{R}(\theta) = \mathbb{E} f_k(\theta) = \mathbb{E}\,\ell(y_k, \langle \theta, \Phi(x_k)\rangle)$:
  - For $0 \le k \le n$, $\mathcal{F}_k = \sigma((x_i, y_i)_{1\le i\le k})$.
  - At step $0 < k \le n$, use a new point independent of θ_{k-1}:

$$f'_k(\theta_{k-1}) = \ell'(y_k, \langle \theta_{k-1}, \Phi(x_k)\rangle)$$

$$\mathbb{E}[f'_k(\theta_{k-1})|\mathcal{F}_{k-1}] = \mathcal{R}'(\theta_{k-1})$$

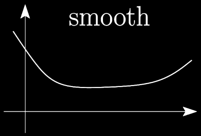
  - Single pass through the data, Running-time $= O(nd)$,
  - "Automatic" regularization.

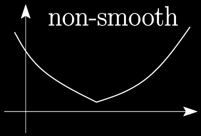
Analysis: Key assumptions: smoothness and/or strong convexity.

Mathematical framework: Smoothness

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

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





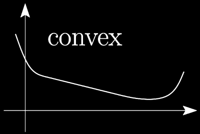
For all $\theta \in \mathbb{R}^d$:

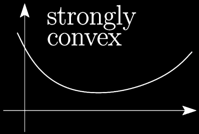
$$g(\theta) \leq g(\theta') + \langle g(\theta'), \theta - \theta' \rangle + L \left\| \theta - \theta' \right\|^2$$

Mathematical framework: Strong Convexity

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

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





For all $\theta \in \mathbb{R}^d$:

$$g(\theta) \geq g(\theta') + \langle g(\theta'), \theta - \theta' \rangle + \mu \left\| \theta - \theta' \right\|^2$$

Application to machine learning

- We consider an a.s. convex loss in θ. Thus $\hat{\mathcal{R}}$ and $\mathcal{R}$ are convex.
- Hessian of $\hat{\mathcal{R}}$ (resp $\mathcal{R}$) $\approx$ covariance matrix $\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\Phi(x_i)^\top$ or $\mathbb{E}[\Phi(X)\Phi(X)^\top]$.

$$\mathcal{R}''(\theta) = \mathbb{E}[\ell''(\langle\theta, \Phi(X)\rangle, Y)\Phi(X)\Phi(X)^\top]$$

- If ℓ is smooth, and $\mathbb{E}[\|\Phi(X)\|^2] \leq r^2$, $\mathcal{R}$ is smooth.
- If ℓ is μ-strongly convex, and data has an invertible covariance matrix (low correlation/dimension), $\mathcal{R}$ is strongly convex.

Analysis: behaviour of $(\theta_n)_{n\geq 0}$

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

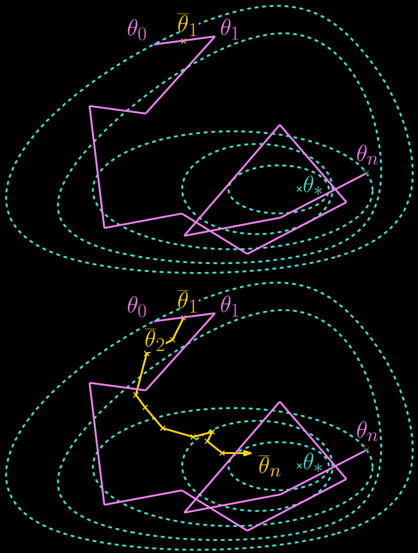
Importance of the learning rate (or sequence of step sizes) $(\gamma_n)_{n\geq 0}$. For smooth and strongly convex problem, traditional analysis shows Fabian (1968); Robbins and Siegmund (1985) that $\theta_n \to \theta_*$ almost surely if

$$\sum_{n=1}^{\infty} \gamma_n = \infty \qquad \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

And asymptotic normality $\sqrt{n}(\theta_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, V)$, for $\gamma_n = \frac{\gamma_0}{n}$, $\gamma_0 \geq \frac{1}{\mu}$.

- Limit variance scales as $1/\mu^2$
- Very sensitive to ill-conditioned problems.
- μ generally unknown, so hard to choose the step size...

Polyak Ruppert averaging



Introduced by Polyak and Juditsky (1992) and Ruppert (1988):

$$\bar{\theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\theta_k.$$

- **off line averaging reduces the noise effect.**
- **on line computing:** $\bar{\theta}_{n+1} = \frac{1}{n+1}\theta_{n+1} + \frac{n}{n+1}\bar{\theta}_n.$
- **one could also consider other averaging schemes (e.g.,**

Convex stochastic approximation: convergence results

- Known **global** minimax rates of convergence for **non-smooth** problems Nemirovsky and Yudin (1983); Agarwal et al. (2012)
  - Strongly convex: $O((\mu n)^{-1})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
  - Non-strongly convex: $O(n^{-1/2})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$
- **Smooth** strongly convex problems
  - All step sizes $\gamma_n = Cn^{-\alpha}$ with $\alpha \in (1/2, 1)$, with averaging, lead to $O(n^{-1})$:
    - asymptotic normality Polyak and Juditsky (1992), with variance independent of μ!
    - non asymptotic analysis Bach and Moulines (2011).
  - Rate $\frac{1}{\mu n}$ for $\gamma_n \propto n^{-1/2}$: adapts to strong convexity.

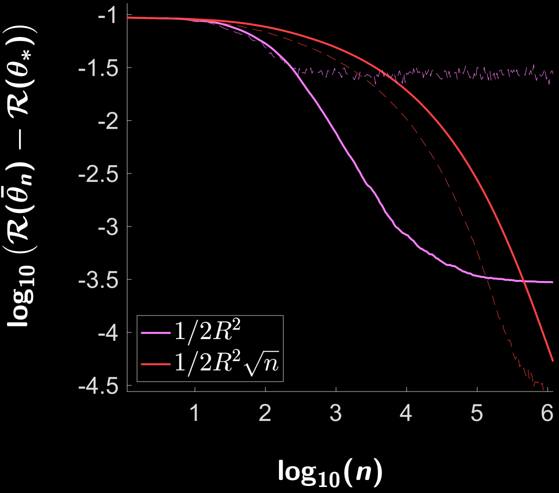
Stochastic Approximation: take home message

- Powerful algorithm:
  - Simple to implement
  - Cheap
  - No regularization needed
- Convergence guarantees:
  - $\gamma_n = \frac{1}{\sqrt{n}}$ good choice in most situations

Problems:

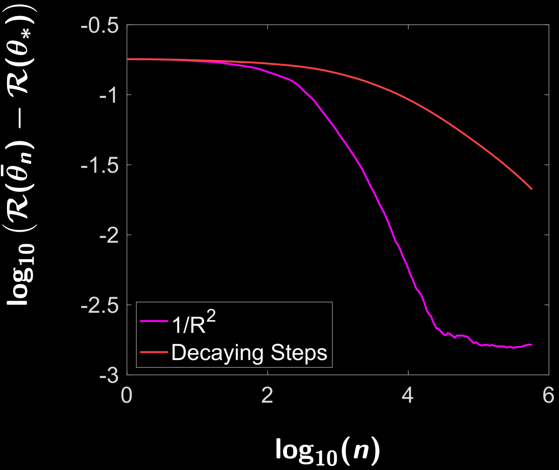
- Initial conditions can be forgotten slowly: could we use even larger step sizes?

Motivation 1/ 2. Large step sizes!



Logistic regression. **Final iterate (dashed), and averaged recursion (plain).**

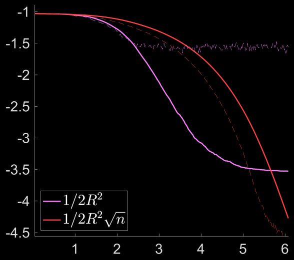
Motivation 1/ 2. Large step sizes, real data



Logistic regression, Covertype dataset, $n = 581012$, $d = 54$.

Comparison between a constant learning rate and decaying learning rate as $\frac{1}{\sqrt{n}}$.

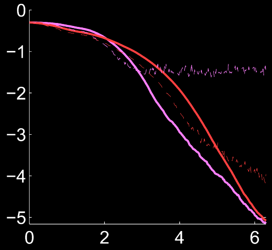
Motivation 2/ 2. Difference between quadratic and logistic loss



Logistic Regression

$$\mathbb{E}\mathcal{R}(\bar{\theta}_n) - \mathcal{R}(\theta_*) = O(\gamma^2)$$

with $\gamma = 1/(2R^2)$



Least-Squares Regression

$$\mathbb{E}\mathcal{R}(\bar{\theta}_n) - \mathcal{R}(\theta_*) = O\left(\frac{1}{n}\right)$$

with $\gamma = 1/(2R^2)$

Larger step sizes: Least-mean-square algorithm

- Least-squares: $\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}\big[(Y - \langle \Phi(X), \theta \rangle)^2\big]$ with $\theta \in \mathbb{R}^d$
  - SGD = least-mean-square algorithm
  - Usually studied without averaging and decreasing step-sizes.
- New analysis for averaging and constant step-size $\gamma = 1/(4R^2)$ Bach and Moulines (2013)
  - 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 the Hessian
  - Main result:

$$\boxed{\mathbb{E}\mathcal{R}(\bar{\theta}_n) - \mathcal{R}(\theta_*) \leqslant \frac{4\sigma^2 d}{n} + \frac{\|\theta_0 - \theta_*\|^2}{\gamma n}}$$

- Matches statistical lower bound Tsybakov (2003).

Related work in Sierra

Led to numerous (non trivial) extensions, at least in our lab !

- **Beyond parametric models: Non Parametric Stochastic Approximation with Large step sizes. Dieuleveut and Bach (2015)**
- **Improved Sampling: Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. Défossez and Bach (2015)**
- **Acceleration: Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. Dieuleveut et al. (2016)**
- **Beyond smoothness and euclidean geometry: Stochastic Composite Least-Squares Regression with convergence rate $O(1/n)$. Flammarion and Bach (2017)**

SGD: an homogeneous Markov chain

Consider a $L-$smooth and $\mu-$strongly convex function $\mathcal{R}$.

SGD with a step-size $\gamma > 0$ is an homogeneous Markov chain:

$$\theta_{k+1}^{\gamma} = \theta_k^{\gamma} - \gamma\left[\mathcal{R}'(\theta_k^{\gamma}) + \varepsilon_{k+1}(\theta_k^{\gamma})\right] ,$$

- satisfies Markov property
- is homogeneous, for γ constant, $(\varepsilon_k)_{k\in\mathbb{N}}$ i.i.d.

Also assume:

- $\mathcal{R}_k' = \mathcal{R}' + \varepsilon_{k+1}$ is almost surely L-co-coercive.
- Bounded moments

$$\mathbb{E}[\|\varepsilon_k(\theta_*)\|^4] < \infty.$$

Stochastic gradient descent as a Markov Chain: Analysis framework†

- Existence of a limit distribution π_γ, and linear convergence to this distribution:

$$\theta_n^\gamma \xrightarrow{d} \pi_\gamma .$$

- Convergence of second order moments of the chain,

$$\bar{\theta}_n^\gamma \xrightarrow[n\to\infty]{L^2} \bar{\theta}_\gamma := \mathbb{E}_{\pi_\gamma}[\theta] .$$

- Behavior under the limit distribution ($\gamma \to 0$): $\bar{\theta}_\gamma = \theta_* + ?$.

↬ Provable convergence improvement with extrapolation tricks.

† Dieuleveut, Durmus, Bach [2017].

Existence of a limit distribution $\gamma \to 0$

Goal: $(\theta_n^\gamma)_{n\geq 0} \xrightarrow{d} \pi_\gamma$.

Theorem

For any $\gamma < (2L)^{-1}$, the chain $(\theta_n^\gamma)_{n\geq 0}$ admits a unique stationary distribution π_γ. In addition for all $\theta_0 \in \mathbb{R}^d$, $n \in \mathbb{N}$:

$$W_2^2(\theta_n^\gamma, \pi_\gamma) \leq (1-\mu\gamma)^n \int_{\mathbb{R}^d} \|\theta_0 - \vartheta\|^2 \, \mathrm{d}\pi_\gamma(\vartheta) \ .$$

Wasserstein metric: distance between probability measures.

Assumptions

A1: f is a μ-strongly convex function.

A2: f is $\mathcal{C}^4$ with bounded second to fourth derivative . Especially, f is L-smooth.

A3: Filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$. For all $k \in \mathbb{N}$, for any $\theta \in \mathbb{R}^d$, $\varepsilon_{k+1}(\theta)$ is an $\mathcal{F}_{k+1}$-measurable random variable and

$$\mathbb{E}\left[\varepsilon_{k+1}(\theta)|\mathcal{F}_k\right] = 0 \ .$$

We assume that the noise functions $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are i.i.d. .

A4: f_1' is almost surely L-co-coercive. Moreover, $\varepsilon_1(\theta_*)$ admits bounded moments up to the order $p \leq 4$:

$$\mathbb{E}^{1/p}[\|\varepsilon_1(\theta_*)\|^p] < \infty.$$

Transition kernel

Fundamental tool: Markov kernel R_γ, (for continuous spaces, $\simeq$ transition matrix in finite state spaces).

Definition

For all initial distributions ν_0 on $\mathcal{B}(\mathbb{R}^d)$ and $k \in \mathbb{N}$, $\nu_0 R_\gamma^k$ denotes the law of θ_k^γ starting at $\theta_0 \sim \nu_0$.

If θ_0 is deterministic, $\theta_k^\gamma \sim \delta_{\theta_0} R_\gamma^k$.

Definition

For any function $h : \mathbb{R}^d \to \mathbb{R}$, $\forall \theta \in \mathbb{R}^d$, $k \geq 1$:

$$R_\gamma^k h(\theta) \;=\; \mathbb{E}_{\theta_0=\theta}[h(\theta_k^\gamma)] = \int_{\mathbb{R}^d} h(\vartheta) \left\{ \delta_\theta R_\gamma^k \right\} (\mathrm{d}\vartheta)$$

notation: for a measure π, function h: $\pi(h) = \int h(\theta) d\pi(\theta)$.

Existence of a limit distribution $\gamma \to 0$

Goal: $(\theta_k^\gamma)_{k\geq 0} \xrightarrow{d} \pi_\gamma$ i.e. $(\nu_0 R_\gamma^k)_{k\geq 0} \to \pi_\gamma$.

Definition

Wasserstein metric: ν and λ probability measures on $\mathbb{R}^d$

$$W_2(\lambda, \nu) := \inf_{\xi \in \Pi(\lambda,\nu)} \left(\int \|x - y\|^2 \xi(dx, dy) \right)^{1/2}$$

$\Pi(\lambda, \nu)$ is the set of probability measure ξ s.t. $A \in \mathcal{B}(\mathbb{R}^d)$, $\xi(A \times \mathbb{R}^d) = \lambda(A)$, $\xi(\mathbb{R}^d \times A) = \nu(A)$.

Theorem

Assume A1:A4, for $\gamma < L^{-1}$, the chain $(\theta_k^\gamma)_{k\geq 0}$ admits a unique stationary distribution π_γ and for all $\theta \in \mathbb{R}^d$, $n \in \mathbb{N}$:

$$W_2^2(\delta_\theta R_\gamma^n, \pi_\gamma) \leq (1 - 2\mu\gamma(1 - \gamma L))^n \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \, \mathrm{d}\pi_\gamma(\vartheta) \ .$$

Existence of a limit distribution: proof I /III

- **Coupling:** θ^1, θ^2 be independent and distributed according to λ_1, λ_2 respectively, and $(\theta^{(1)}_{k,\gamma})_{\geq 0}$,$(\theta^{(2)}_{k,\gamma})_{k\geq 0}$ SGD iterates:

$$\begin{cases} \theta^{(1)}_{k+1,\gamma} & = \theta^{(1)}_{k,\gamma} - \gamma\big[f'(\theta^{(1)}_{k,\gamma}) + \varepsilon_{k+1}(\theta^{(1)}_{k,\gamma})\big] \\ \theta^{(2)}_{k+1,\gamma} & = \theta^{(2)}_{k,\gamma} - \gamma\big[f'(\theta^{(2)}_{k,\gamma}) + \varepsilon_{k+1}(\theta^{(2)}_{k,\gamma})\big] \end{cases} .$$

- for all $k \geq 0$, the distribution of $(\theta^{(1)}_{k,\gamma}, \theta^{(2)}_{k,\gamma})$ is in $\Pi(\lambda_1 R^k_\gamma, \lambda_2 R^k_\gamma)$

Existence of a limit distribution: proof II/III

$$
\begin{aligned}
W_2^2(\lambda_1 R_\gamma, \lambda_2 R_\gamma) &\leq \mathbb{E}\left[\|\theta^{(1)}_{1,\gamma} - \theta^{(2)}_{1,\gamma}\|^2\right] \\
&\leq \mathbb{E}\left[\|\theta^1 - \gamma f_1'(\theta^1) - (\theta^2 - \gamma f_1'(\theta^2)))\|^2\right] \\
&\overset{A3}{\leq} \mathbb{E}\left[\left\|\theta^1 - \theta^2\right\|^2 - 2\gamma\left\langle f'(\theta^1) - f'(\theta^2), \theta^1 - \theta\right.\right. \\
&\quad + \gamma^2 \mathbb{E}\left[\left\|f_1'(\theta^1) - f_1'(\theta^2)\right\|^2\right] \\
&\overset{A4}{\leq} \mathbb{E}\left[\left\|\theta^1 - \theta^2\right\|^2\right] \\
&\quad -2\gamma(1-\gamma L)\left\langle f'(\theta^1) - f'(\theta^2), \theta^1 - \theta^2\right\rangle \\
&\overset{A1}{\leq} (1 - 2\mu\gamma(1-\gamma L))\mathbb{E}\left[\left\|\theta^1 - \theta^2\right\|^2\right] ,
\end{aligned}
$$

define $\rho = (1 - 2\mu\gamma(1 - \gamma L))$.

Existence of a limit distribution: proof III/III

By induction:

$$W_2^2(\lambda_1 R_\gamma^n, \lambda_2 R_\gamma^n) \leq \mathbb{E}\left[\|\theta_{n,\gamma}^{(1)} - \theta_{n,\gamma}^{(2)}\|^2\right] \leq \rho^n \int_{x,y} \|x - y\|^2 \,\mathrm{d}\lambda_1(x)\mathrm{d}$$

- Thus $W_2(\delta_x R_\gamma^n, \delta_y R_\gamma^n) \leq (1 - 2\mu\gamma(1 - \gamma L))^n \|x - y\|^2$.
- { prob. measures with second order moment }: Polish space.
- Picard fixed point theorem, $(\lambda_1 R_\gamma^n)_{n \geq 0}$ is a Cauchy sequence and converges to a limit $\pi_\gamma^{\lambda_1}$.
- Uniqueness, invariance, and Theorem follow:

$$W_2^2(\delta_\theta R_\gamma^n, \pi_\gamma) \leq (1 - 2\mu\gamma(1 - \gamma L))^n \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \,\mathrm{d}\pi_\gamma(\vartheta) .$$

Consequence: solutions to the Poisson equation.

In the following, we will need to introduce, for any ϕ sufficiently regular (say L_ϕ-Lipshitz) a function ψ_ϕ s.t., for $\theta \in \mathbb{R}^d$:

$$\psi_\phi(\theta) \;=\; \sum_{k=0}^{\infty} \left(\mathbb{E}_{\theta_0=\theta}\left[\phi(\theta_k^\gamma)\right] - \mathbb{E}_{\pi_\gamma}(\phi(\theta)) \right)$$

As $\left| \mathbb{E}_{\theta_0=\theta}\left[\phi(\theta_k^\gamma)\right] - \mathbb{E}_{\pi_\gamma}(\phi(\theta)) \right| \leq L_\phi W_2(\delta_\theta R_\gamma^k, \pi_\gamma)$, the sum absolutely converges for all θ. Moreover, ψ is also Lipshitz, and satisfies:

$$(I - R_\gamma)\psi = \phi - \pi_\gamma(\phi).$$

Which is the "Poisson Equation".

Behavior under limit distribution.

Ergodic theorem: $\bar{\theta}_n \to \mathbb{E}_{\pi_\gamma}[\theta] =: \bar{\theta}_\gamma$. Where is $\bar{\theta}_\gamma$?

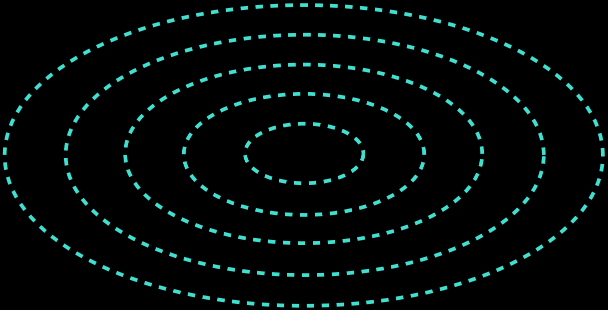
If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

$$\theta_1^\gamma = \theta_0^\gamma - \gamma\big[\mathcal{R}'(\theta_0^\gamma) + \varepsilon_1(\theta_0^\gamma)\big] \,.$$

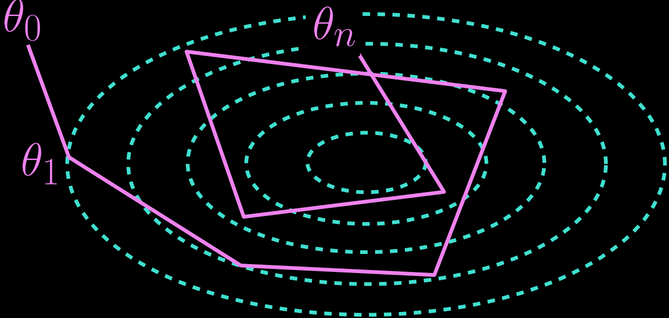
$$\mathbb{E}_{\pi_\gamma}\left[\mathcal{R}'(\theta)\right] = 0$$

In the quadratic case (linear gradients) $\Sigma\mathbb{E}_{\pi_\gamma}[\theta - \theta_*] = 0$:
$\bar{\theta}_\gamma = \theta_*$!

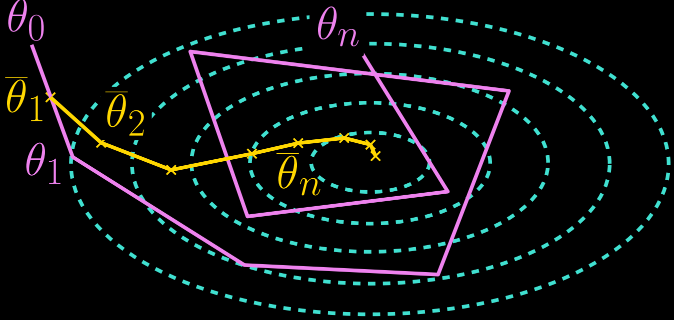
Constant learning rate SGD: convergence in the quadratic case



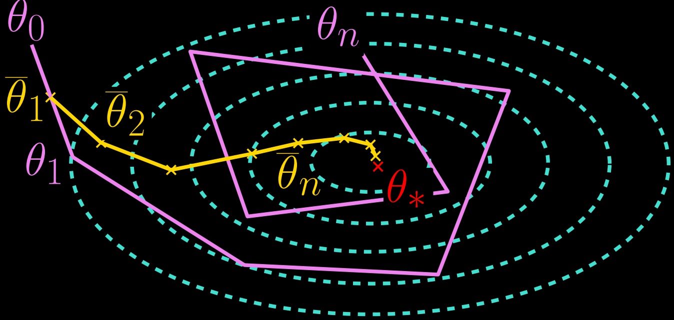
Constant learning rate SGD: convergence in the quadratic case



Constant learning rate SGD: convergence in the quadratic case



Constant learning rate SGD: convergence in the quadratic case



Behavior under limit distribution.

Ergodic theorem: $\bar{\theta}_n \to \mathbb{E}_{\pi_\gamma}[\theta] =: \bar{\theta}_\gamma$. Where is $\bar{\theta}_\gamma$?

If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

$$\theta_1^\gamma = \theta_0^\gamma - \gamma\left[\mathcal{R}'(\theta_0^\gamma) + \varepsilon_1(\theta_0^\gamma)\right] .$$

$$\mathbb{E}_{\pi_\gamma}\left[\mathcal{R}'(\theta)\right] = 0$$

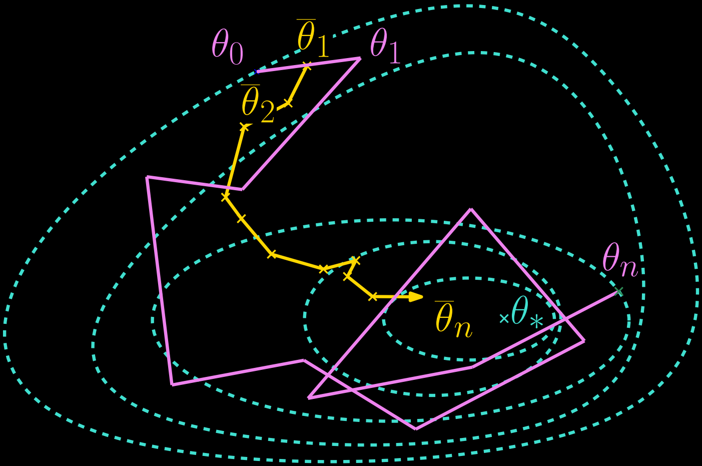
In the quadratic case (linear gradients) $\Sigma\mathbb{E}_{\pi_\gamma}[\theta - \theta_*] = 0$:
$\bar{\theta}_\gamma = \theta_*$!

In the general case, Taylor expansion of $\mathcal{R}$, and same reasoning on higher moments of the chain leads to

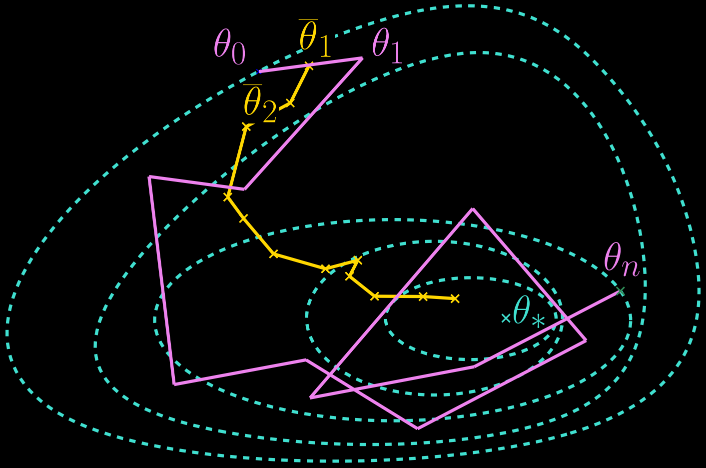
$$\bar{\theta}_\gamma - \theta_* \simeq \gamma\mathcal{R}''(\theta_*)^{-1}\mathcal{R}'''(\theta_*)\Big(\big[\mathcal{R}''(\theta_*) \otimes I + I \otimes \mathcal{R}''(\theta_*)\big]^{-1}\mathbb{E}_\varepsilon[\varepsilon(\theta_*)^{\otimes 2}]\Big)$$

Overall, $\bar{\theta}_\gamma - \theta_* = \gamma\Delta + O(\gamma^2)$.

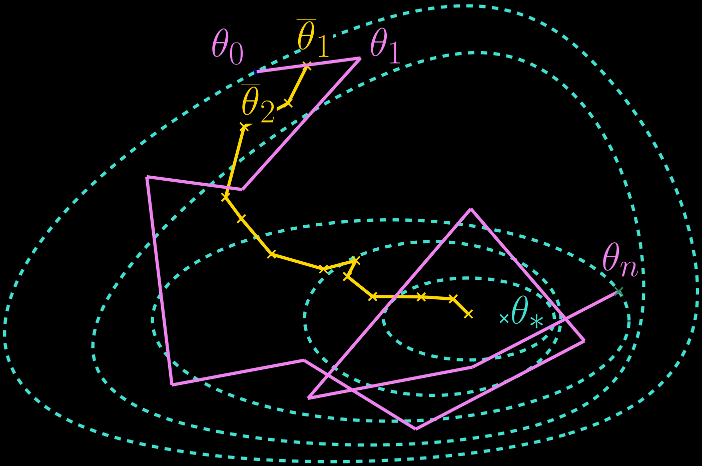
Constant learning rate SGD: convergence in the non-quadratic case



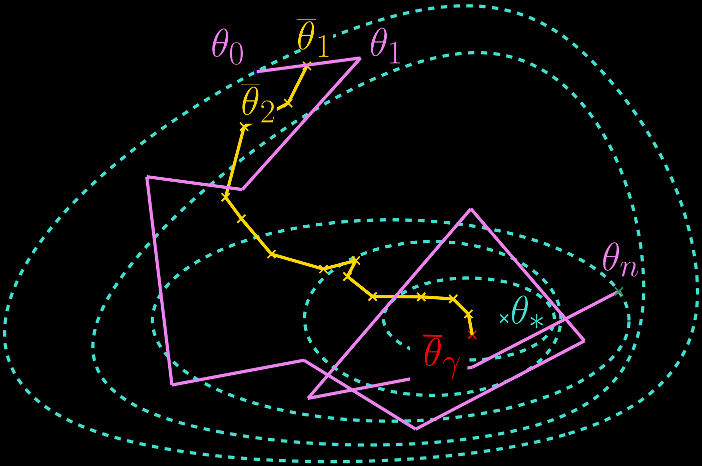
Constant learning rate SGD: convergence in the non-quadratic case



Constant learning rate SGD: convergence in the non-quadratic case



Constant learning rate SGD: convergence in the non-quadratic case



Convergence of second order moments, $\gamma > 0$, $n \to +\infty$.

Non asymptotic bound for the convergence $\bar{\theta}_n^\gamma - \theta_*$:

Proposition (Convergence of the Markov chain)

Let $\gamma \in]0, 1/(2L)[$ and assume A1-A4. With $\rho := (1 - \gamma\mu)^{1/2}$:

$$
\begin{aligned}
\mathbb{E}\bar{\theta}_k^\gamma - \bar{\theta}_\gamma &= \frac{1}{k}\int_{\mathbb{R}^d} \psi_\gamma(\theta)\mathrm{d}\nu_0(\theta) + O(\rho^k)\ , \\
\mathbb{E}\left[(\bar{\theta}_k^\gamma - \bar{\theta}_\gamma)^{\otimes 2}\right] &= \frac{1}{k}\int_{\mathbb{R}^d} \left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma - \varphi)(\theta)(\psi_\gamma - \varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) \\
&\quad + \frac{1}{k^2}\int_{\mathbb{R}^d} \left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top + \chi_\gamma^1(\theta) - \chi_\gamma^2(\theta)\right]\mathrm{d}\nu_0(\theta) + O(\rho^k)\ .
\end{aligned}
$$

- $\phi(\theta) = \theta - \theta_*$. ψ_γ Poisson solution associated to ϕ,
- χ_γ^1 Poisson solution associated to $\phi\phi^\top$,
- χ_γ^2 Poisson solution associated to $(\psi_\gamma - \phi)(\psi_\gamma - \phi)^\top$.

Bias - Variance decomposition.

Convergence of second order moments, proof.

- Algebraic calculation (R_γ encodes a linear relationship between the distributions of θ_k^γ)
- For the first result:

$$
\begin{aligned}
\mathbb{E}\left[\bar{\theta}_k^\gamma\right] - \theta_* &= \frac{1}{k}\sum_{i=0}^{k-1}(R_\gamma^i \varphi)(\theta_0) \\
&= \pi_\gamma \varphi + \frac{1}{k}\psi_\gamma(\theta_0) + R_\gamma^k \psi_\gamma(\theta_0)
\end{aligned}
$$

using $R_\gamma^i \pi_\gamma(\varphi) = \pi_\gamma \varphi$, and $R_\gamma^k \psi_\gamma(\theta_0) = O(\rho^k)$

Recovering Least mean squares

If $f(\theta) = \frac{1}{2}\mathbb{E}_\rho\left[(Y - \langle \Phi(X), \theta \rangle)^2\right]$, then we can compute the Poisson solutions: recovers Défossez and Bach (2015).

Corollary (Convergence in the quadratic case)

Consider LMS with $\gamma L \leq 1/2$, and denoting ξ the additive part of the noise*, one has:

$$\mathbb{E}\left[(\bar\theta_k^\gamma - \theta_*)^{\otimes 2}\right] = \frac{1}{k^2\gamma^2}\Sigma^{-1}\Omega(\theta_0 - \theta_*)^{\otimes 2}\Sigma^{-1} + \frac{1}{k}\Sigma^{-1}\left[\mathbb{E}\varepsilon^{\otimes 2}\right]\Sigma^{-1} - \frac{1}{k^2\gamma}\Sigma^{-1}\Omega\left[\Sigma \otimes I + I \otimes \Sigma - \gamma T\right]^{-1}\left[\mathbb{E}\xi^{\otimes 2}\right]\Sigma^{-1} + O(\rho^k)$$

with $\Omega := (\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)(\Sigma \otimes I + I \otimes \Sigma - \gamma T)^{-1}$, and $T : A \mapsto \mathbb{E}\left[(x^\top A x) x x^\top\right]$.

$$\mathbb{E}\left[(\bar\theta_k^\gamma - \theta_*)^{\otimes 2}\right] \simeq \underbrace{\frac{1}{k^2\gamma^2}\Sigma^{-1}(\theta_0 - \theta_*)^{\otimes 2}\Sigma^{-1}}_{\text{Bias}} + \underbrace{\frac{1}{k}\Sigma^{-1}\left[\mathbb{E}\varepsilon^{\otimes 2}\right]\Sigma^{-1}}_{\text{Variance}} + O(\rho^k)\ .$$

$^* f_n'(\theta) = (\Phi(x_n)\Phi(x_n)^\top - \Sigma)(\theta - \theta_*) + (\langle \theta_*, \Phi(x_n)\rangle - y_n)\Phi(x_n)$

Take home message

- Convergence in distribution of the MC (Wasserstein metric).
- Allows to prove and analyze convergence of the moments of the chain to 0 (can be generalized to any function).
- We provide second order development as $\gamma \to 0$:

$$\bar{\theta}_\gamma = \theta_* + \gamma \Delta_1 + \gamma^2 \Delta_2 + o(\gamma^2).$$

- Error decomposition as a sum of three terms :

$$f(\bar{\theta}_n^\gamma) - f(\theta_*) \leq \frac{Bias}{\gamma^2 n^2 \mu} + \frac{Var}{n} + \frac{\gamma^2}{\mu},$$

- As a consequence, we can recover the rate, for $\gamma = 1/\sqrt{n}$:

$$f(\bar{\theta}_n^\gamma) - f(\theta_*) = O\left(\frac{1}{n\mu}\right).$$

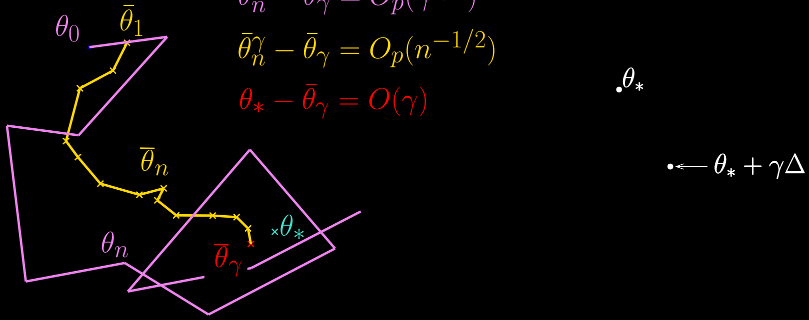
- Beyond: comparison to the continuous gradient flow for a more general approach.

Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



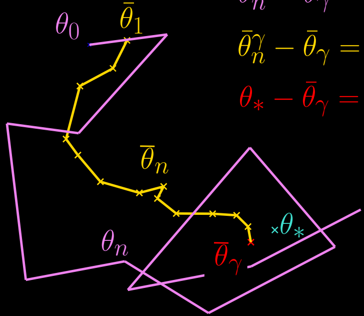
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

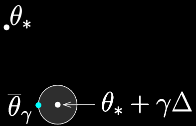
Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$





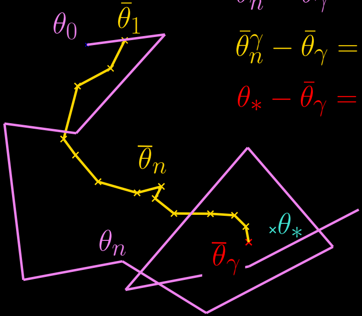
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

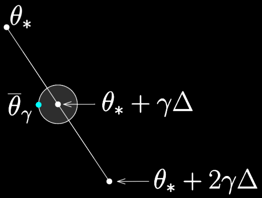
Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$





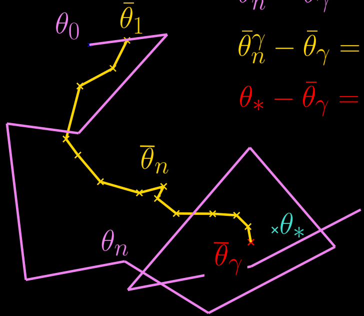
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

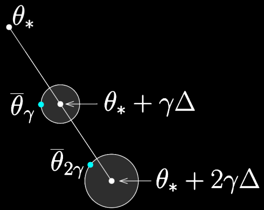
Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$





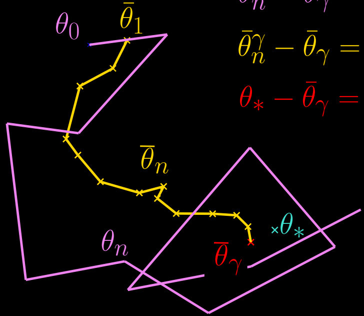
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

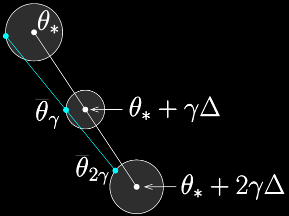
Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

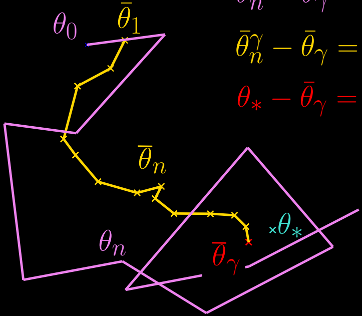
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$





Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

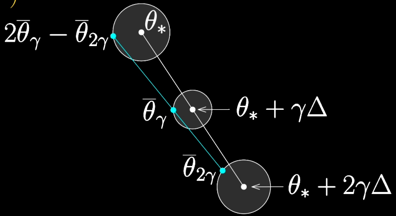
Richardson extrapolation



$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

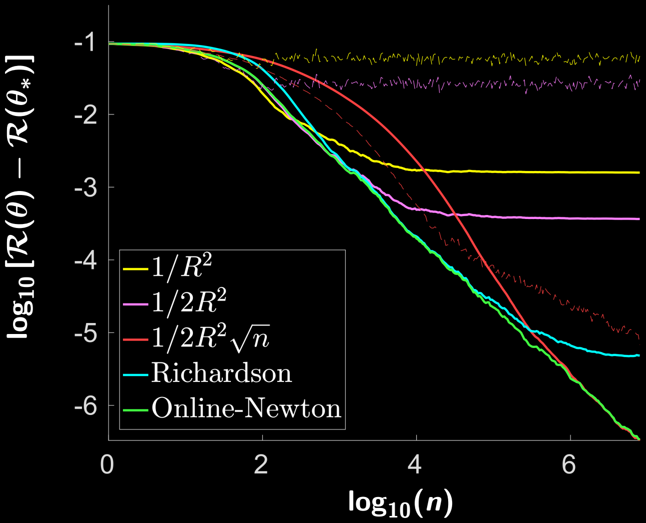
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



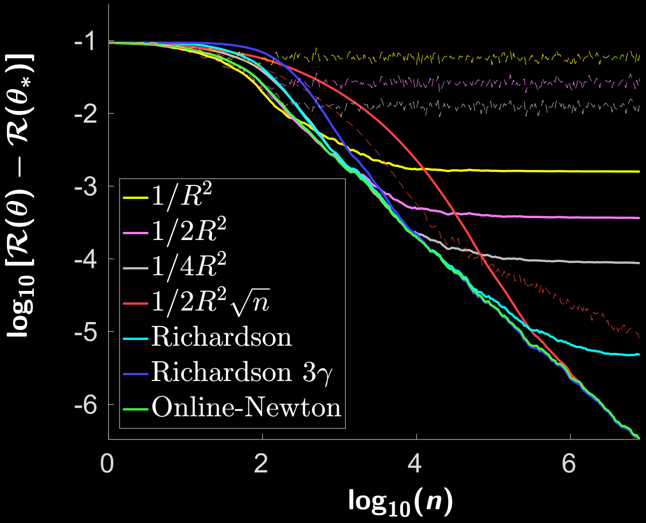
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

Experiments



Synthetic data, logistic regression, $n = 8.10^6$

Experiments: Double Richardson



Synthetic data, logistic regression, $n = 8.10^6$

"Richardson 3γ": estimator built using Richardson on 3 different sequences: $\tilde{\theta}_n^3 = \frac{8}{3}\bar{\theta}_n^{\gamma} - 2\bar{\theta}_n^{2\gamma} + \frac{1}{3}\bar{\theta}_n^{4\gamma}$

Real data

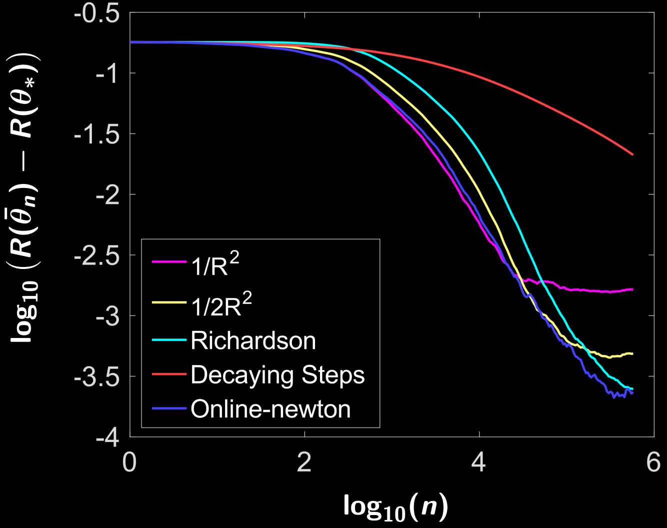


Figure 1: Logistic regression, Covertype dataset. $n = 581012$, $d = 54$.

Directions

Open directions:

- Extending proofs to self-concordant setting.
- Does this three term decomposition extend to decaying steps.
- Understand the convex case more precisely.

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. Ann. Statist., 40(5):2452–2482.

Bach, F. and Moulines, E. (2011). Non-asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS'11, pages 451–459, USA. Curran Associates Inc.

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

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

Défossez, A. and Bach, F. (2015). Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS).

Dieuleveut, A. and Bach, F. (2015). Non-parametric stochastic approximation with large step sizes. Annals of Statistics.

Dieuleveut, A., Flammarion, N., and Bach, F. (2016). Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. ArXiv e-prints.

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

Flammarion, N. and Bach, F. (2017). Stochastic composite least-squares regression with convergence rate o (1/n).

Jones, G. L. (2004). On the Markov chain central limit theorem. Probability Surveys, 1:299–320.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an O(1/t) rate for the stochastic projected subgradient method. ArXiv e-prints 1212.2002.

Nemirovsky, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

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

Robbins, H. and Monro, S. (1951). A stochastic approxiation method. The Annals of mathematical Statistics, 22(3):400–407.

Robbins, H. and Siegmund, D. (1985). A convergence theorem for non negative almost supermartingales and some applications. In Herbert Robbins Selected Papers, pages 111–135. Springer.

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

Tsybakov, A. B. (2003). Optimal rates of aggregation. In Proceedings of the Annual Conference on Computational Learning Theory.