Bridging the gap between Stochastic Approximation and Markov chains

Aymeric DIEULEVEUT

ENS Paris, INRIA

July 11, 2017

Joint work with Francis Bach and Alain Durmus.

Supervised Machine Learning

Input/output pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{X} = \mathbb{R}^d$, following some unknown distribution ρ.

$\mathcal{Y} = \mathbb{R}$ (regression) or $\{-1, 1\}$ (classification).

Goal: find a function $\theta : \mathcal{X} \to \mathbb{R}$, such that $\langle \theta, \Phi(X) \rangle$ is close to Y, for some features $\Phi(X) \in \mathbb{R}^d$.

Loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$: squared loss, logistic loss, 0-1 loss, etc.

Supervised Machine Learning

Input/output pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{X} = \mathbb{R}^d$, following some unknown distribution ρ.

$\mathcal{Y} = \mathbb{R}$ (regression) or $\{-1, 1\}$ (classification).

Goal: find a function $\theta : \mathcal{X} \to \mathbb{R}$, such that $\langle \theta, \Phi(X) \rangle$ is close to Y, for some features $\Phi(X) \in \mathbb{R}^d$.

Loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$: squared loss, logistic loss, 0-1 loss, etc.

Risk (or generalization error) as

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

Minimization problem:

$$\theta_* = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta)$$

Stochastic Approximation Framework

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

Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

$\mathbb{E}[f_n'(\theta_{n-1})|\mathcal{F}_{n-1}] = f'(\theta_{n-1})$ for a filtration $(\mathcal{F}_n)_{n\geq 0}$, θ_n is $\mathcal{F}_n$ measurable.

Stochastic Approximation Framework

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

Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

$\mathbb{E}[f_n'(\theta_{n-1})|\mathcal{F}_{n-1}] = f'(\theta_{n-1})$ for a filtration $(\mathcal{F}_n)_{n\geq 0}$, θ_n is $\mathcal{F}_n$ measurable.

Polyak-Ruppert averaging considers:

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

Stochastic Approximation in Machine learning

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

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

For the risk $R(\theta) = \mathbb{E} f_k(\theta) = \mathbb{E}\, \ell(y_k, \langle \theta, \Phi(x_k) \rangle)$:

Stochastic Approximation in Machine learning

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

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

For the risk $R(\theta) = \mathbb{E} f_k(\theta) = \mathbb{E}\, \ell(y_k, \langle \theta, \Phi(x_k) \rangle)$:

For $0 \leq k \leq n$, $\mathcal{F}_k = \sigma((x_i, y_i)_{1 \leq i \leq k})$.

At step $0 < k \leq n$, use a **new point** independent of θ_{k-1}:

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

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

Stochastic Approximation in Machine learning

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

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

For the risk $R(\theta) = \mathbb{E} f_k(\theta) = \mathbb{E}\, \ell(y_k, \langle \theta, \Phi(x_k) \rangle)$:

For $0 \leq k \leq n$, $\mathcal{F}_k = \sigma((x_i, y_i)_{1 \leq i \leq k})$.

At step $0 < k \leq n$, use a **new point** independent of θ_{k-1}:

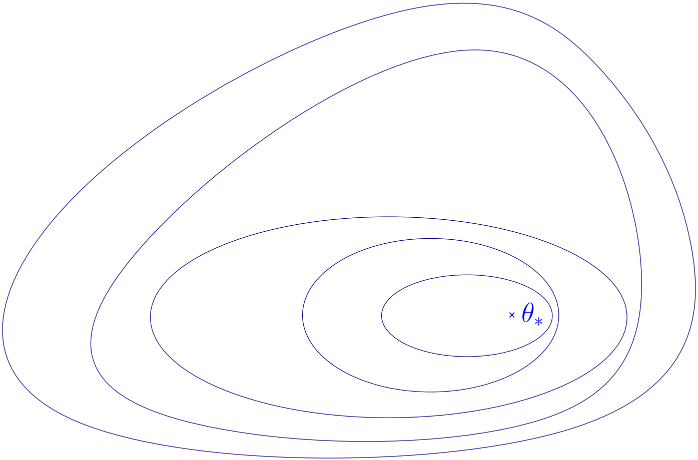
$$f_k'(\theta_{k-1}) = \ell'(y_k, \langle \theta_{k-1}, \Phi(x_k) \rangle)$$

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

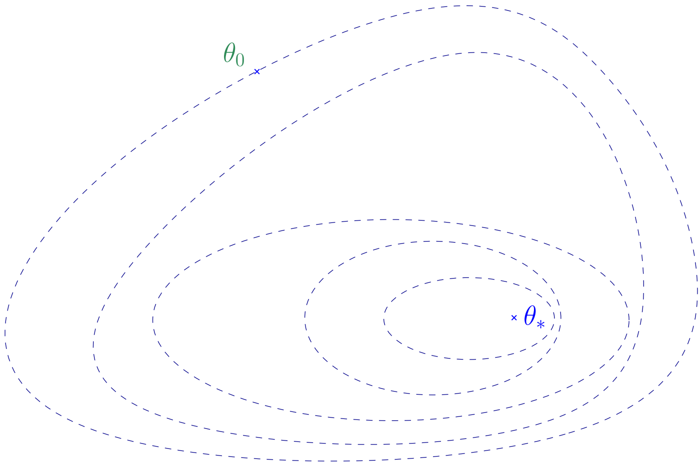
Single pass through the data

"Automatic" regularization.

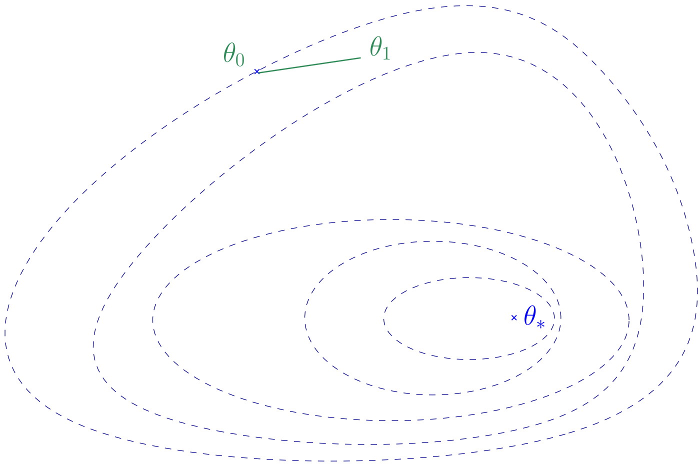
Stochastic gradient descent



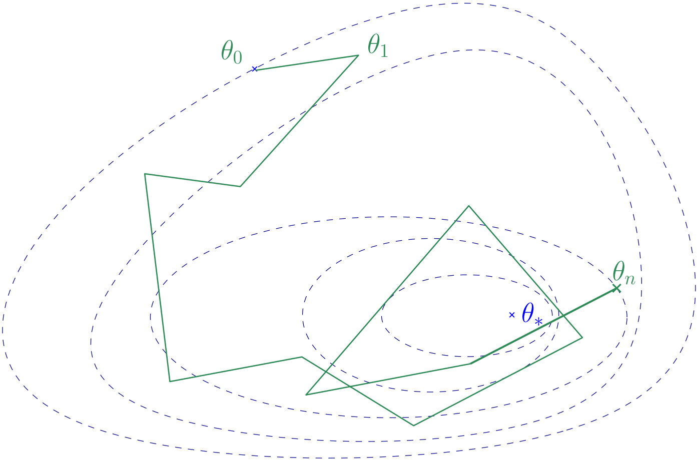
Stochastic gradient descent



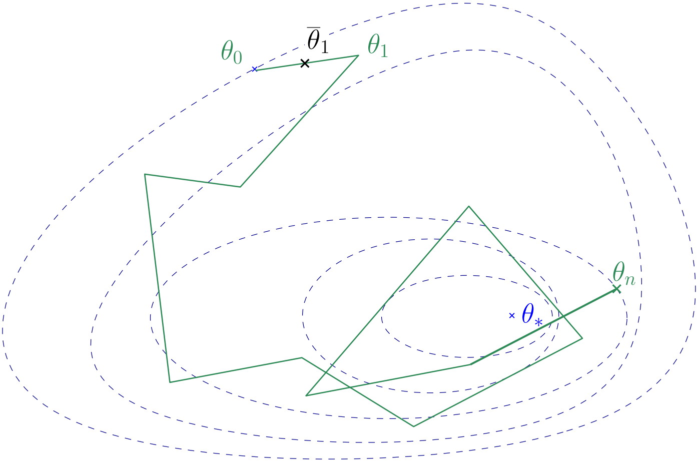
Stochastic gradient descent



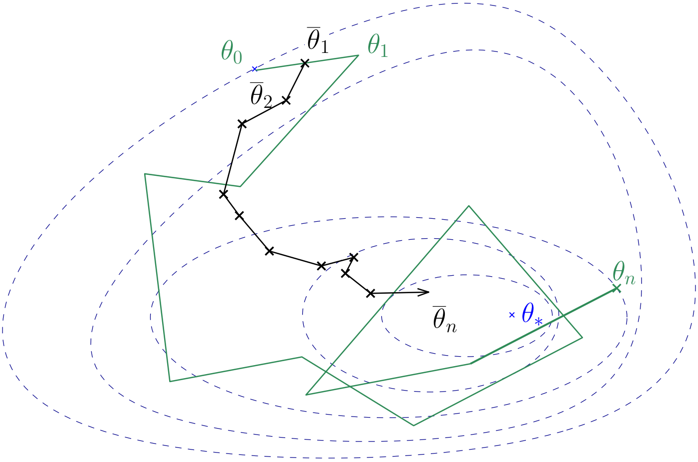
Stochastic gradient descent



Stochastic gradient descent: averaging !



Stochastic gradient descent: averaging !



Convex stochastic approximation: convergence results

Smooth Non-strongly convex: $O(n^{-1/2})$
Attained by **averaged** stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

Smooth μ-strongly convex problems: $O(\frac{1}{\mu n})$
also for $\gamma_n \propto n^{-1/2}$:
$\looparrowright$ **adapts to strong convexity.**

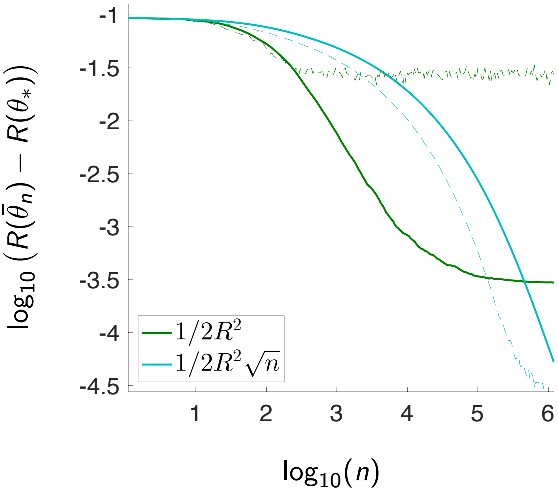


Figure 1: Logistic regression (smooth strongly convex[2] problem), dimension 25. Comparison between a constant learning rate and decaying learning rate as $\frac{1}{\sqrt{n}}$. Final iterate (dashed), and averaged recursion (plain)

[1]in fact, only self concordant but behaves similarly. [Bach, 2014]

Real data

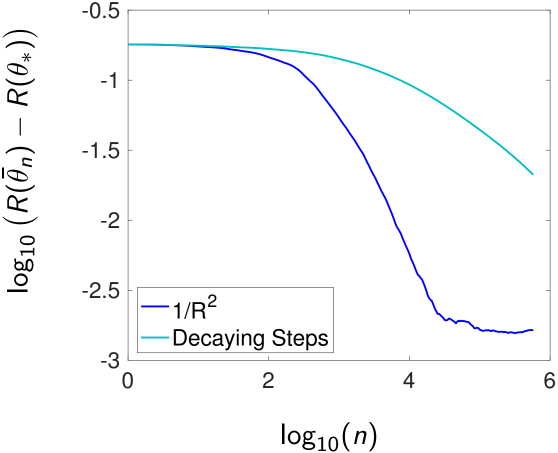
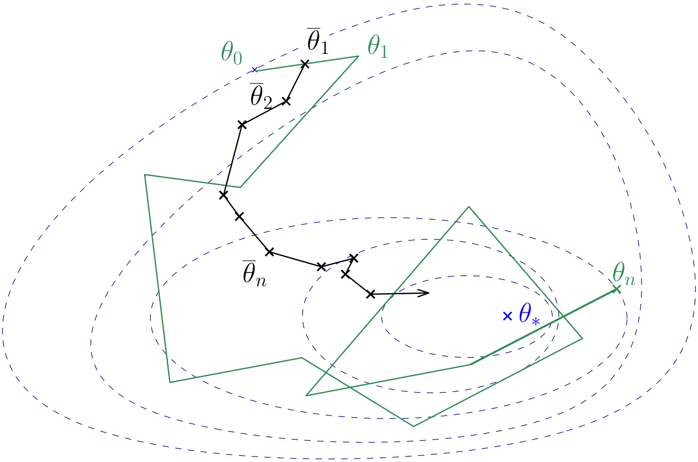


Figure 2: Logistic regression, Covertype dataset, $n = 581012$, $d = 54$. Comparison between a constant learning rate and decaying learning rate as $\frac{1}{\sqrt{n}}$.

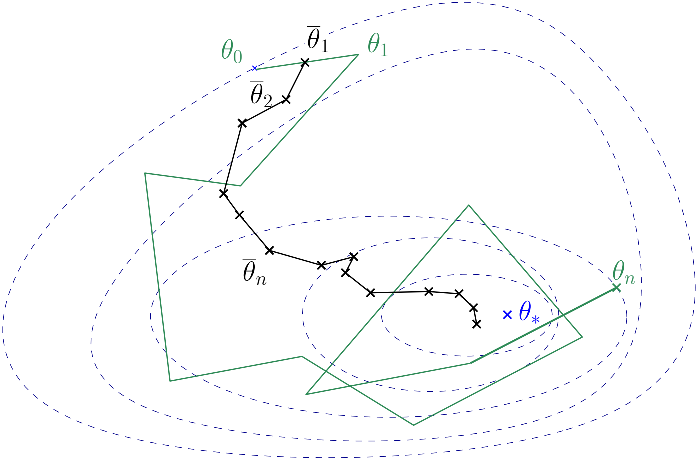
Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



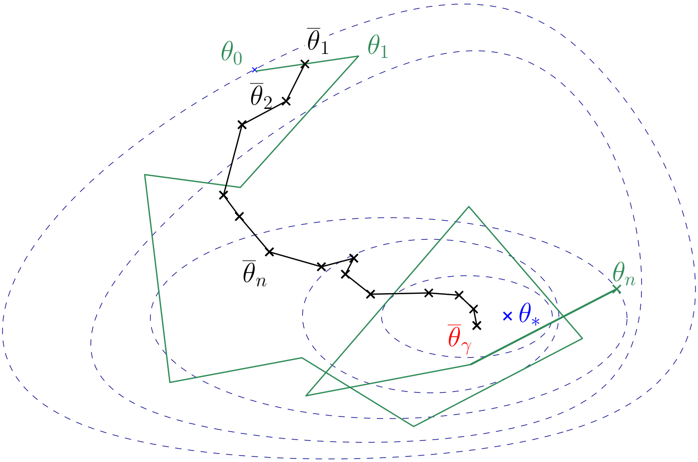
Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



Behavior under limit distribution.

π_γ limit distribution of (θ_n).

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

Behavior under limit distribution.

π_γ limit distribution of (θ_n).

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

$\theta_{1,\gamma} = \theta_{0,\gamma} - \gamma\big[f'(\theta_{0,\gamma}) + \varepsilon_1(\theta_{0,\gamma})\big]$. If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

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

Behavior under limit distribution.

π_γ limit distribution of (θ_n).

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

$\theta_{1,\gamma} = \theta_{0,\gamma} - \gamma\big[f'(\theta_{0,\gamma}) + \varepsilon_1(\theta_{0,\gamma})\big]$. If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

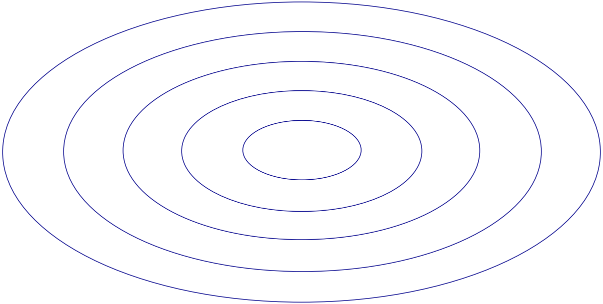
$$\mathbb{E}_{\pi_\gamma}\left[f'(\theta)\right] = 0$$

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

$$\bar{\theta}_\gamma = \theta_*!$$

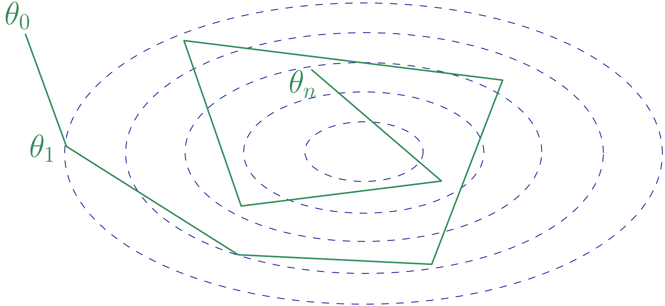
Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



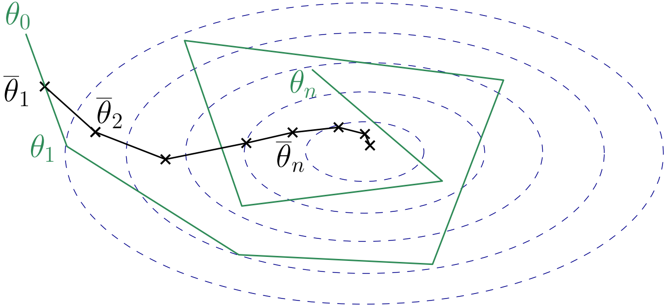
Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



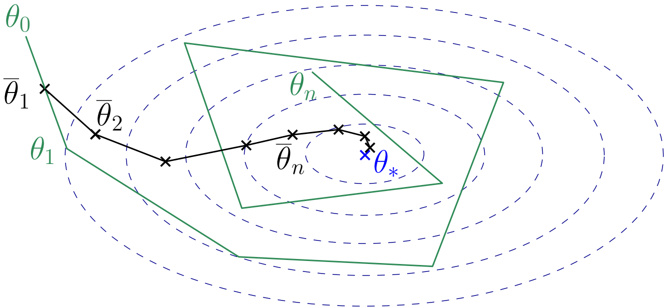
Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



Constant learning rate SGD: a Markov Chain

If $\gamma = C$ (possibly $C(n)$), then SGD is an *homogeneous Markov chain*.



Ergodic theorem:

$$\sqrt{n}(\bar{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1} C \Sigma^{-1})$$

with Σ covariance matrix, $C = \mathbb{E}[(Y - \langle \Phi(X), \theta_* \rangle)^2 \Phi(X)\Phi(X)^\top]$.

Convex stochastic approximation: convergence results

Smooth Non-strongly convex: $O(n^{-1/2})$
Attained by **averaged** stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

Smooth strongly convex problems $O(\frac{1}{\mu n})$
also for $\gamma_n \propto n^{-1/2}$:
↬ **adapts to strong convexity.**

Least-squares

Averaging and constant step-size $\gamma = 1/(4R^2)$
[Bach and Moulines, 2013]

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

Matches statistical lower bound [Tsybakov, 2003].

Behavior under limit distribution.

π_γ limit distribution of (θ_n).

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

$\theta_{1,\gamma} = \theta_{0,\gamma} - \gamma\big[f'(\theta_{0,\gamma}) + \varepsilon_1(\theta_{0,\gamma})\big]$. If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

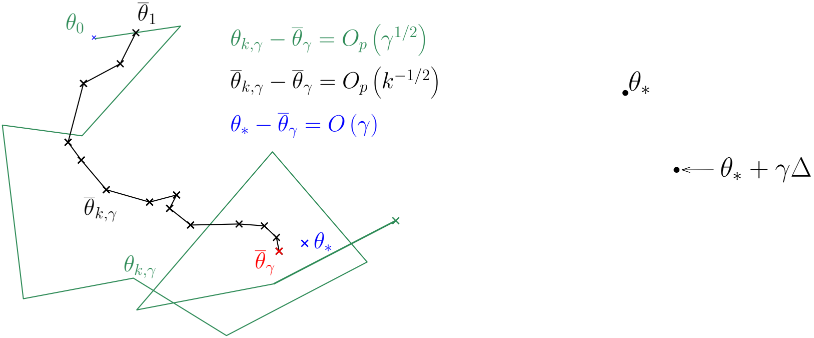
$$\mathbb{E}_{\pi_\gamma}\left[f'(\theta)\right] = 0$$

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

$$\bar{\theta}_\gamma = \theta_*!$$

In the general case, $\bar{\theta}_\gamma - \theta_* = \gamma\Delta_1 + \gamma^2\Delta_2 + o(\gamma^2)$.

What we have :

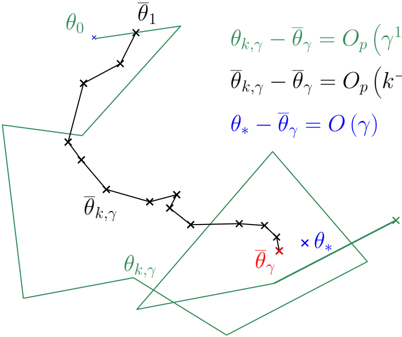


$$\theta_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(\gamma^{1/2}\right)$$

$$\overline{\theta}_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(k^{-1/2}\right)$$

$$\theta_* - \overline{\theta}_\gamma = O\left(\gamma\right)$$

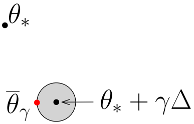
What we have :



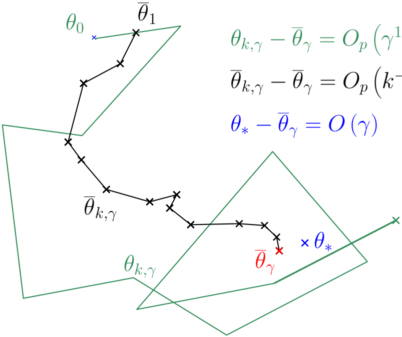
$$\theta_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(\gamma^{1/2}\right)$$

$$\overline{\theta}_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(k^{-1/2}\right)$$

$$\theta_* - \overline{\theta}_\gamma = O\left(\gamma\right)$$



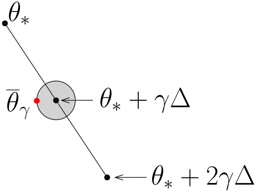
What we have :



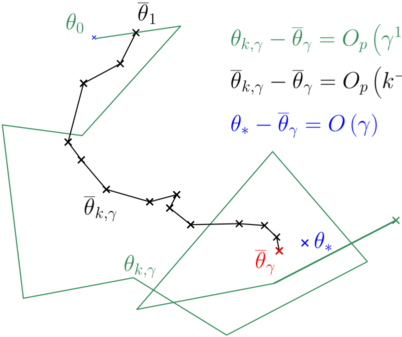
$$\theta_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(\gamma^{1/2}\right)$$

$$\overline{\theta}_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(k^{-1/2}\right)$$

$$\theta_* - \overline{\theta}_\gamma = O\left(\gamma\right)$$



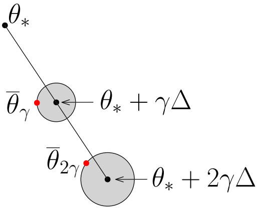
What we have :



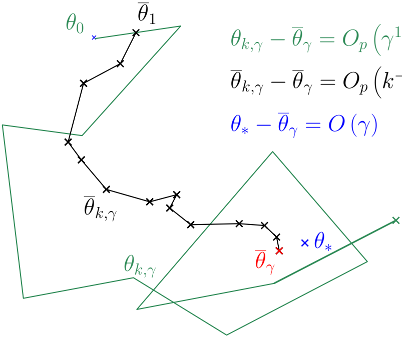
$$\theta_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(\gamma^{1/2}\right)$$

$$\overline{\theta}_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(k^{-1/2}\right)$$

$$\theta_* - \overline{\theta}_\gamma = O\left(\gamma\right)$$



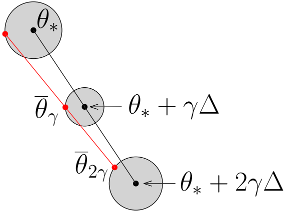
What we have :



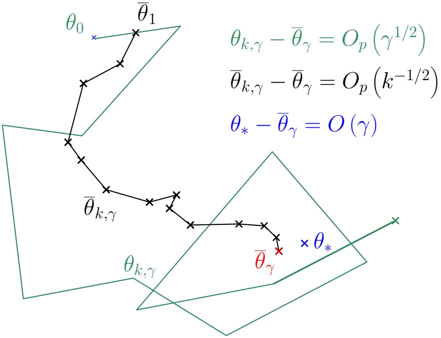
$$\theta_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(\gamma^{1/2}\right)$$

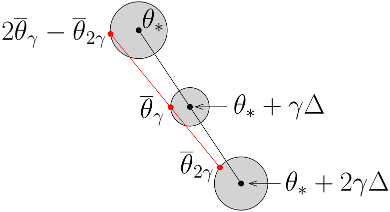
$$\overline{\theta}_{k,\gamma} - \overline{\theta}_\gamma = O_p\left(k^{-1/2}\right)$$

$$\theta_* - \overline{\theta}_\gamma = O\left(\gamma\right)$$



What we have :





Recovering convergence closer to θ_* by **Richardson extrapolation**

$$2\bar{\theta}_{n,\gamma} - \bar{\theta}_{n,2\gamma}$$

Experiments

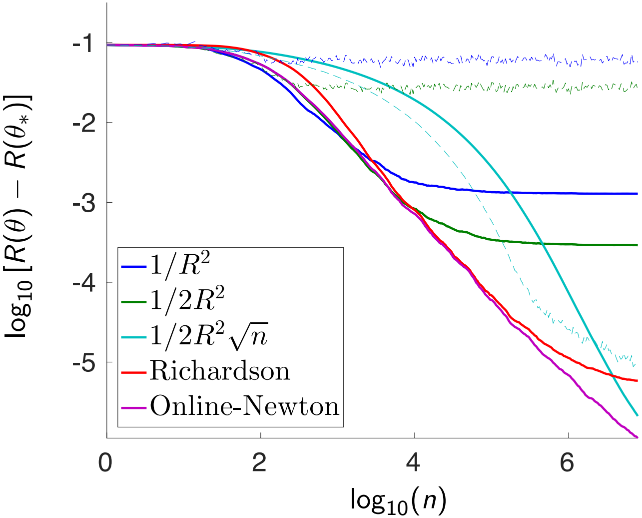


Figure 3: Synthetic data, logistic regression, $d = 12$, $n = 8.10^6$, averaged over 50 repetitions.

Experiments: Double Richardson

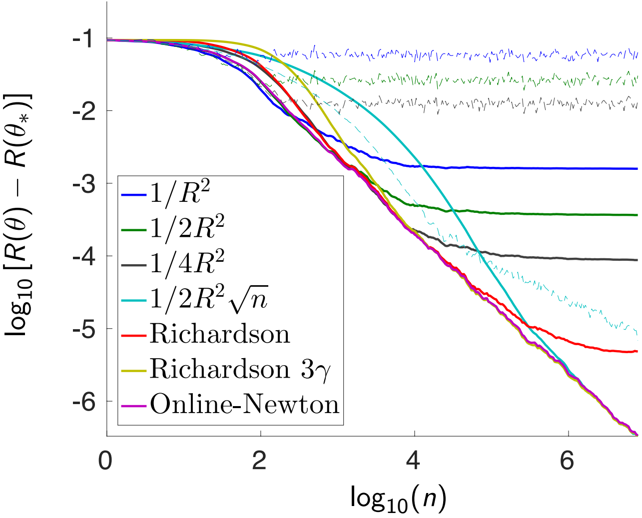


Figure 4: Synthetic data, logistic regression, $d = 4$, $n = 8.10^6$, averaged over 50 repetitions. "Richardson 3γ": estimator built using *Richardson on 3 different sequences*: $\tilde{\theta}_n^3 = \frac{8}{3}\bar{\theta}_{n,\gamma} - 2\bar{\theta}_{n,2\gamma} + \frac{1}{3}\bar{\theta}_{n,4\gamma}$

Stochastic gradient descent as a Markov Chain: Analysis framework

Analysis outline:

Existence of a limit distribution π_γ, and fast convergence to this distribution.

Behavior under the limit distribution ($\gamma \to 0$),

Convergence of second order moments of the chain ($n \to \infty$),

Recovering LMS,

Comparison to the gradient flow.

Richardson-Romberg iteration

Soon online.

Stochastic gradient descent as a Markov Chain: Analysis framework

Analysis outline:

Existence of a limit distribution π_γ, and fast convergence to this distribution.

Behavior under the limit distribution ($\gamma \to 0$),

Convergence of second order moments of the chain ($n \to \infty$),

Recovering LMS,

Comparison to the gradient flow.

Richardson-Romberg iteration

Soon online.

Assumptions

f is a μ-strongly convex function, L-smooth.

Unbiased gradients

$$\mathbb{E}\left[f'_{k+1}(\theta)|\mathcal{F}_k\right] = f'(\theta_k)\ .$$

f_k a.s. L-smooth, and convex. It implies, $\forall\theta, \eta$

$$\left\|f_1'(\theta) - f_1'(\eta)\right\|^2 \leq L\Big\langle f'(\theta) - f'(\eta), \theta - \eta\Big\rangle$$

Existence of a limit distribution: proof I /III

If $\theta_0 \sim \lambda_1$ then

$$\theta_{k,\gamma} \sim \lambda_1 R_\gamma^k$$

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:

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

Existence of a limit distribution: proof II/III

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

Bach, F. (2014).
Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression.
J. Mach. Learn. Res., 15(1):595–627.

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).

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

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