Stochastic Algorithms in Machine Learning

Aymeric DIEULEVEUT

EPFL, Lausanne

December 1st, 2017

Journée Algorithmes Stochastiques, Paris Dauphine

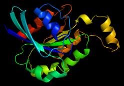
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Outline

1. Machine learning context.
2. Stochastic algorithms to minimize Empirical Risk .
3. Stochastic Approximation: using stochastic gradient descent (SGD) to minimize Generalization Risk.
4. Markov chain: insightful point of view on constant step size Stochastic Approximation.

Supervised Machine Learning: definition & applications

Goal: predict a phenomenon from "explanatory variables", given a set of observations.



Bio-informatics

Input: DNA/RNA sequence,
Output: Disease predisposition / Drug responsiveness
$n \to 10$ to 10^4
d (e.g., number of basis) $\to 10^6$

Image classification

Input: Handwritten digits / Images,
Output: Digit
$n \to$ up to 10^9
d (e.g., number of pixels) $\to 10^6$

"Large scale" learning framework: both the number of examples n and the number of explanatory variables d are large.

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).
- Goal: 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.
- Define the Generalization risk (a.k.a., generalization error, "true risk") as

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

Empirical Risk minimization (I)

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, **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 (ERM) (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 $\hat{\theta}$.

Take home

- Problem is formalized as a (convex) optimization problem.
- In the large scale setting, **high dimensional problem** and **many examples**.

Stochastic algorithms for ERM

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

1. High dimension $d \implies$ First order algorithms

Gradient Descent (GD) :

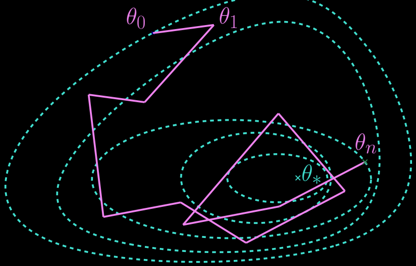
$$\boxed{\theta_k = \theta_{k-1} - \gamma_k \hat{\mathcal{R}}'(\theta_{k-1})}$$

Problem: computing the gradient costs $O(dn)$ per iteration.

2. Large $n \implies$ Stochastic algorithms

Stochastic Gradient Descent (SGD)

Stochastic Gradient descent



- **Goal**:

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

  given unbiased gradient estimates f'_n
- $\theta_* := \mathrm{argmin}_{\mathbb{R}^d} f(\theta)$.
- **Key algorithm: Stochastic Gradient Descent (SGD)** (Robbins and Monro, 1951):

$$\boxed{\theta_k = \theta_{k-1} - \gamma_k f'_k(\theta_{k-1})}$$

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

SGD for ERM: $f = \hat{\mathcal{R}}$

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

$$f_j(\theta) := \ell(y_j, \langle \theta, \Phi(x_j) \rangle).$$

One observation at each step $\implies$ complexity $O(d)$ per iteration.

For the empirical risk $\hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(y_k, \langle \theta, \Phi(x_k) \rangle)$.

- At each step $k \in \mathbb{N}^*$, sample $I_k \sim \mathcal{U}\{1, \ldots n\}$, and use:

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

- with $\mathcal{F}_k = \sigma((x_i, y_i)_{1 \leq i \leq n}, (I_i)_{1 \leq i \leq k})$,

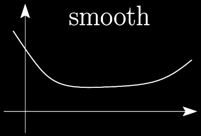
$$\mathbb{E}[f'_{I_k}(\theta_{k-1}) | \mathcal{F}_{k-1}] = \frac{1}{n} \sum_{k=1}^{n} \ell'(y_k, \langle \theta, \Phi(x_k) \rangle) = \hat{\mathcal{R}}'(\theta_{k-1}).$$

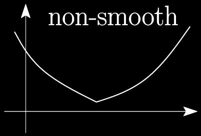
Mathematical framework: smoothness and/or strong convexity.

Mathematical framework: Smoothness

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

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





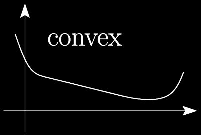
For all $\theta \in \mathbb{R}^d$:

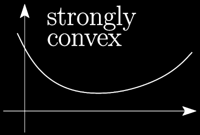
$$g(\theta) \leq g(\theta') + \langle g(\theta'), \theta - \theta' \rangle + L \left\| \theta - \theta' \right\|^2$$

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 $\hat{\mathcal{R}} \approx$ covariance matrix $\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top$ $(\simeq \mathbb{E}[\Phi(X)\Phi(X)^\top].)$

$$\hat{\mathcal{R}}''(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\ell''(\langle\theta, \Phi(X_i)\rangle, Y_i)\Phi(x_i)\Phi(x_i)^\top\right)$$

- 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 $(\theta_n)_{n\geq 0}$

$$\boxed{\theta_k = \theta_{k-1} - \gamma_k f_k'(\theta_{k-1})}$$

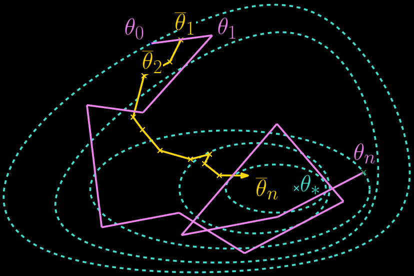
Importance of the **learning rate** (or sequence of step sizes) $(\gamma_k)_{k\geq 0}$. For smooth and strongly convex problem, traditional analysis shows Fabian (1968); Robbins and Siegmund (1985) that $\theta_k \to \theta_*$ almost surely if

$$\sum_{k=1}^{\infty} \gamma_k = \infty \qquad \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

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

- 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):

$$\bar{\theta}_k = \frac{1}{k+1}\sum_{i=0}^{k}\theta_i.$$

- off line averaging reduces the noise effect.
- on line computing: $\bar{\theta}_{k+1} = \frac{1}{k+1}\theta_{k+1} + \frac{k}{k+1}\bar{\theta}_k$.
- one could also consider other averaging schemes (e.g., Lacoste-Julien et al. (2012)).

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: $O((\mu k)^{-1})$
    Attained by **averaged** stochastic gradient descent with $\gamma_k \propto (\mu k)^{-1}$
  - Non-strongly convex: $O(k^{-1/2})$
    Attained by **averaged** stochastic gradient descent with $\gamma_k \propto k^{-1/2}$
- **Smooth strongly convex problems**
  - Rate $\frac{1}{\mu k}$ for $\gamma_k \propto k^{-1/2}$: adapts to strong convexity.

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

(all rates have hidden dependences in the smoothness)

	$\min \hat{\mathcal{R}}$	
	SGD	GD
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k}\right)$
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\mu k})$

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

(all rates have hidden dependences in the smoothness)

$\min \hat{\mathcal{R}}$

	SGD	GD
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k}\right)$
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\mu k})$

⊖ Gradient descent update costs n times as much as SGD update.

Can we get best of both worlds ?

Methods for finite sum minimization

Key idea: using a random gradient with less variance.

- **GD**: at step k, use $\frac{1}{n}\sum_{i=0}^{n} f_i'(\theta_k)$
- **SGD**: at step k, sample $i_k \sim \mathcal{U}[1;n]$, use $f_{i_k}'(\theta_k)$
- **SAG**: at step k,
  - keep a "full gradient" $\frac{1}{n}\sum_{i=0}^{n} f_i'(\theta_{k_i})$, with $\theta_{k_i} \in \{\theta_1, \dots \theta_k\}$
  - sample $i_k \sim \mathcal{U}[1;n]$, use

$$\frac{1}{n}\left(\sum_{i=0}^{n} f_i'(\theta_{k_i}) - f_{i_k}'(\theta_{k_{i_k}}) + f_{i_k}'(\theta_k)\right),$$

$\looparrowright$ $\oplus$ update costs the same as SGD

$\looparrowright$ $\ominus$ needs to store all gradients $f_i'(\theta_{k_i})$ at "points in the past"

Some references:

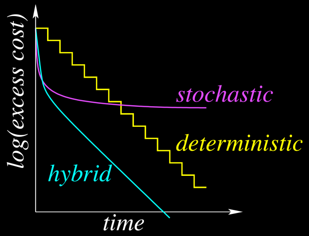
- SAG Schmidt et al. (2013), SAGA Defazio et al. (2014a)
- SVRG Johnson and Zhang (2013) (reduces memory cost but 2 epochs...)
- FINITO Defazio et al. (2014b)
- S2GD Konečný and Richtárik (2013)...

And many others... See for example Niao He's lecture notes for a nice overview.

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

$\min \hat{\mathcal{R}}$

	SGD	GD	SAG
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k}\right)$	
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\mu k})$	$O\left(1 - (\mu \wedge \frac{1}{n})\right)^k$



GD, SGD, SAG (Fig. from Schmidt et al. (2013))

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

	$\min \hat{\mathcal{R}}$		
	SGD	GD	SAG
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k}\right)$	
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\mu k})$	$O\left(1 - (\mu \wedge \frac{1}{n})\right)^k$
Lower Bounds	α	β	γ

α : Stoch. opt. information theoretic lower bounds, Agarwal et al. (2012);

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

	$\min \hat{\mathcal{R}}$		
	SGD	AGD	SAG
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k^2}\right)$	
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\sqrt{\mu}k})$	$O\left(1 - (\mu \wedge \frac{1}{n})\right)^k$
Lower Bounds	α	β	γ

α : Stoch. opt. information theoretic lower bounds, Agarwal et al. (2012);
β: Black box first order optimization, Nesterov (2004);
γ: Lower bounds for optimizing finite sums, Agarwal and Bottou (2014).

Take home

Stochastic algorithms for Empirical Risk Minimization.

- Several algorithms to optimize empirical risk, most efficient ones are stochastic and rely on finite sum structure
- Stochastic algorithms to optimize a deterministic function.
- Rates depend on the regularity of the function.

What about generalization risk

Generalization guarantees:

- Uniform upper bound $\sup_\theta \left| \hat{\mathcal{R}}(\theta) - \mathcal{R}(\theta) \right|$. (empirical process theory)
- More precise: localized complexities (Bartlett et al., 2002), stability (Bousquet and Elisseeff, 2002).

Problems for ERM:

- Choose regularization (overfitting risk)
- How many iterations (i.e., passes on the data)?
- Generalization guarantees generally of order $O(1/\sqrt{n})$, no need to be precise

2 important insights:

1. No need to optimize below statistical error,
2. Generalization risk is more important than empirical risk.

SGD can be used to minimize the generalization risk.

SGD for the generalization risk: $f = \mathcal{R}$

SGD: key assumption $\mathbb{E}[f_n'(\theta_{n-1})|\mathcal{F}_{n-1}] = f'(\theta_{n-1})$.

For the risk

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

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

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

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

$$\begin{aligned}\mathbb{E}[f_k'(\theta_{k-1})|\mathcal{F}_{k-1}] &= \mathbb{E}_\rho[\ell'(y_k, \langle\theta_{k-1}, \Phi(x_k)\rangle)|\mathcal{F}_{k-1}] \\ &= \mathbb{E}_\rho\left[\ell'(Y, \langle\theta_{k-1}, \Phi(X)\rangle)\right] = \mathcal{R}'(\theta_{k-1})\end{aligned}$$

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

	ERM minimization	Gen. risk minimization
	several passes : $0 \leq k$	One pass $0 \leq k \leq n$
x_i, y_i is	$\mathcal{F}_t$-measurable for any t	$\mathcal{F}_t$-measurable for $t \geq i$.

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

	$\min \hat{\mathcal{R}}$			$\min \mathcal{R}$
	SGD	AGD	SAG	SGD
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k^2}\right)$		$O\left(\frac{1}{\sqrt{k}}\right)$
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\sqrt{\mu}k})$	$O\left(1 - (\mu \wedge \frac{1}{n})\right)^k$	$O\left(\frac{1}{\mu k}\right)$

Convergence rate for $f(\tilde{\theta}_k) - f(\theta_*)$, **smooth** objective f.

	$\min \hat{\mathcal{R}}$			$\min \mathcal{R}$
	SGD	AGD	SAG	SGD
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k^2}\right)$		$O\left(\frac{1}{\sqrt{n}}\right)$
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\sqrt{\mu}k})$	$O\left(1-(\mu \wedge \frac{1}{n})\right)^k$	$O\left(\frac{1}{\mu n}\right)$
		$0 \leq k$		$0 \leq k \leq n$
Lower Bounds	α	β	γ	δ

δ : Information theoretic LB - Statistical theory (Tsybakov, 2003).

Gradient is unknown

Least Mean Squares: rate independent of μ

- **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).
- Optimal rate with "large" (constant) step sizes

Take home

- SGD can be used to minimize the true risk directly
- Stochastic algorithm to minimize unknown function
- No regularization needed, only one pass
- For Least Squares, with constant step, optimal rate .

Take home

- SGD can be used to minimize the true risk directly
- Stochastic algorithm to minimize unknown function
- No regularization needed, only one pass
- For Least Squares, with constant step, optimal rate .

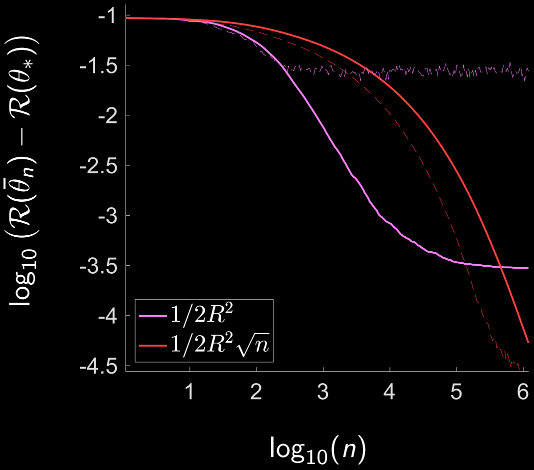
$\looparrowright$**Stochastic approximation, beyond Least Squares ?**

Beyond finite dimensional Least squares

- 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)
- General smooth and strongly convex optimization: Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains (Dieuleveut et al., 2017).

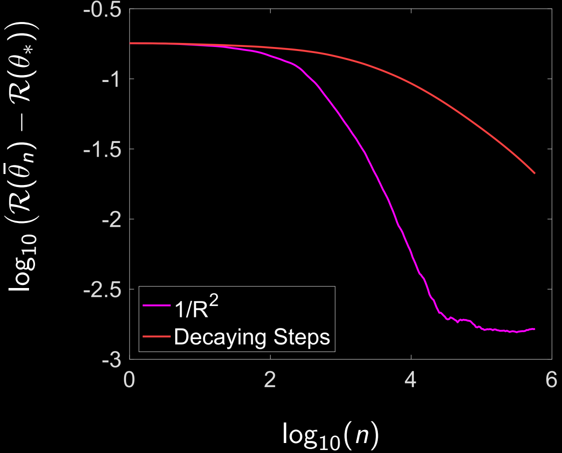
Beyond least squares. Logistic regression

$$\min_{\theta \in \mathbb{R}^d} \quad \mathbb{E} \log \Big(1 + \exp(-Y \langle \theta, \Phi(X) \rangle) \Big).$$



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

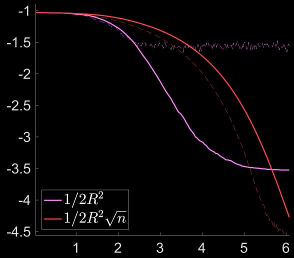
Beyond least squares. Logistic regression, real data



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

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

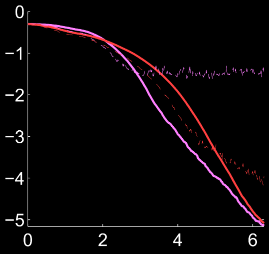
Motivation 2/ 2. Difference between quadratic and logistic loss



Logistic Regression

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

with $\gamma = 1/(4R^2)$



Least-Squares Regression

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

with $\gamma = 1/(4R^2)$

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:

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

- 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:

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

- Convergence of second order moments of the chain,

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

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

$\looparrowright$ Provable convergence improvement with extrapolation tricks.

† *Dieuleveut, Durmus, Bach [2017]*.

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

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

Theorem

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

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

Wasserstein metric: distance between probability measures.

Behavior under limit distribution.

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

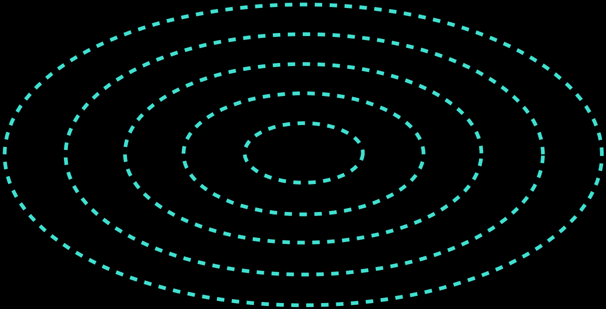
If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

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

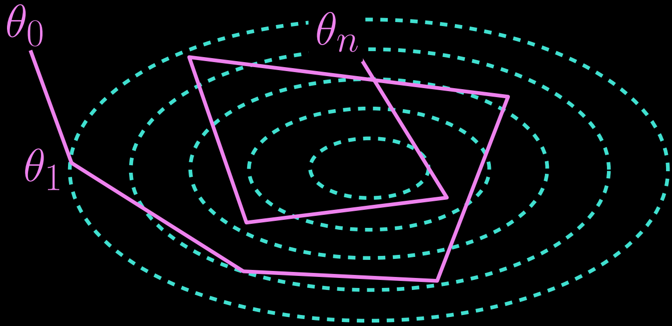
$$\mathbb{E}_{\pi_\gamma}\left[\mathcal{R}'(\theta)\right] = 0$$

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

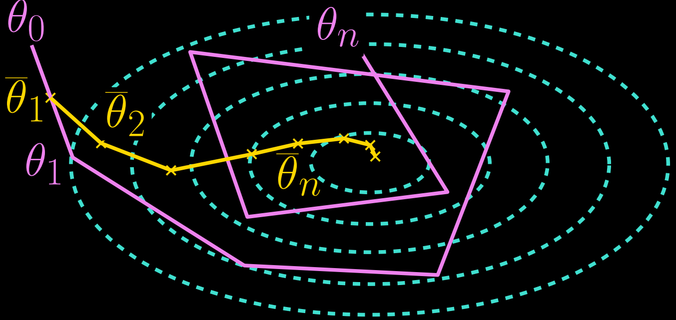
Constant learning rate SGD: convergence in the quadratic case



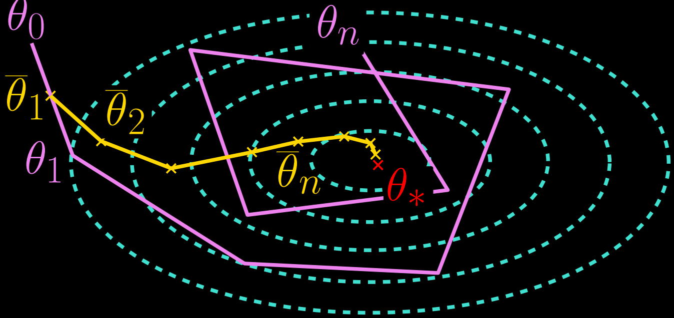
Constant learning rate SGD: convergence in the quadratic case



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 $\theta_1 \sim \pi_\gamma$.

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

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

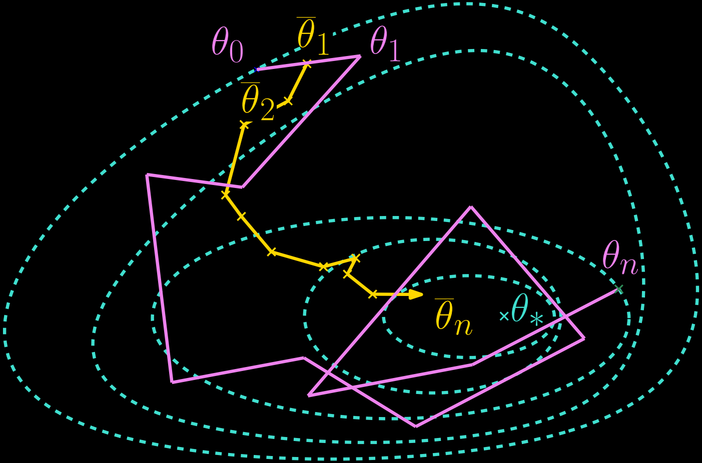
In the quadratic case (linear gradients) $\Sigma\mathbb{E}_{\pi_\gamma}[\theta - \theta_*] = 0$: $\bar{\theta}_\gamma = \theta_*$!

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

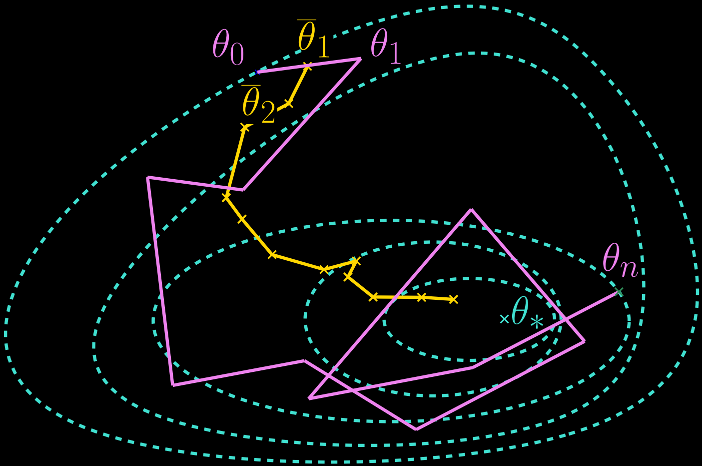
$$\bar{\theta}_\gamma - \theta_* = \gamma\mathcal{R}''(\theta_*)^{-1}\mathcal{R}'''(\theta_*)\Big(\left[\mathcal{R}''(\theta_*) \otimes I + I \otimes \mathcal{R}''(\theta_*)\right]^{-1}\mathbb{E}_\varepsilon[\varepsilon(\theta_*)^{\otimes 2}]\Big) + O(\gamma^2)$$

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

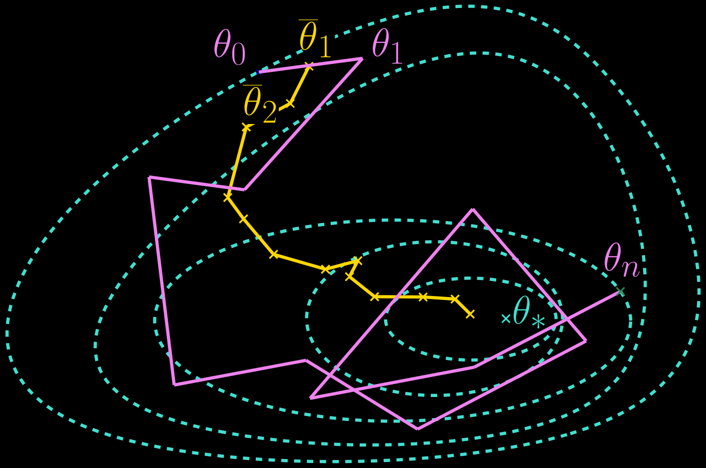
Constant learning rate SGD: convergence in the non-quadratic case



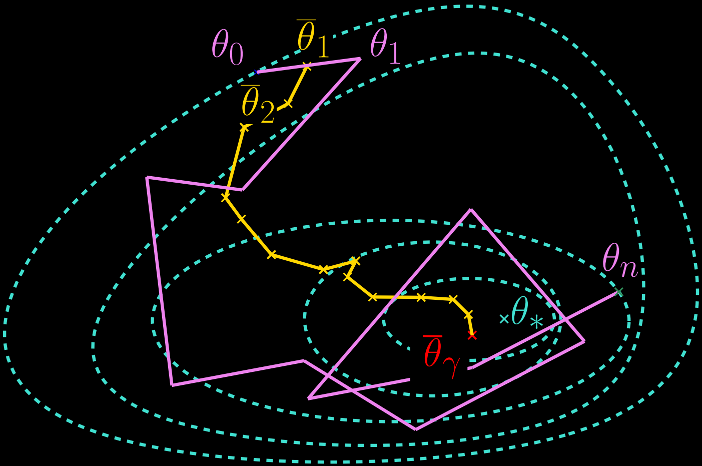
Constant learning rate SGD: convergence in the non-quadratic case



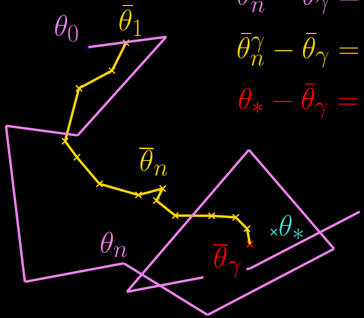
Constant learning rate SGD: convergence in the non-quadratic case



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



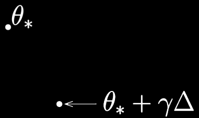
Richardson extrapolation



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

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

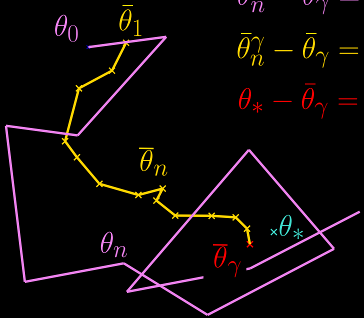
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



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

$$2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$$

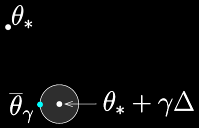
Richardson extrapolation



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

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

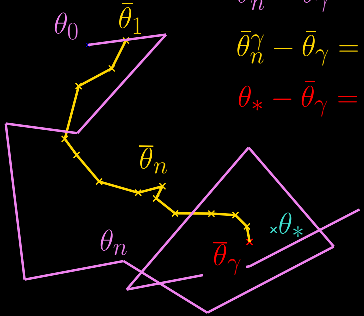
$$\theta_* - \bar\theta_\gamma = O(\gamma)$$



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

$$2\bar\theta_n^\gamma - \bar\theta_n^{2\gamma}$$

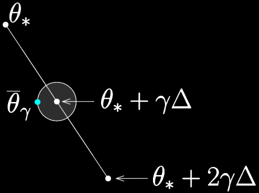
Richardson extrapolation



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

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

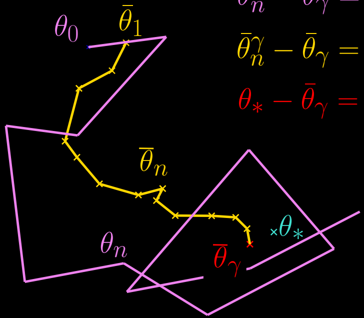
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



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

$$2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$$

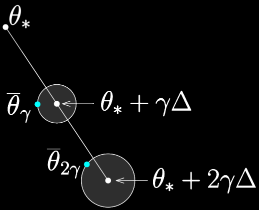
Richardson extrapolation



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

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

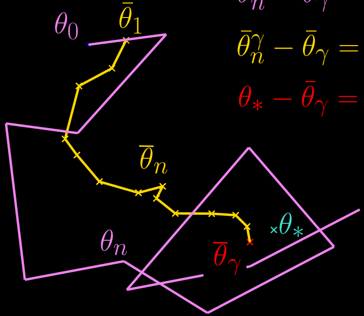
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



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

$$2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$$

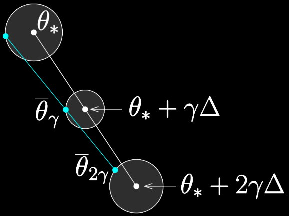
Richardson extrapolation



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

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

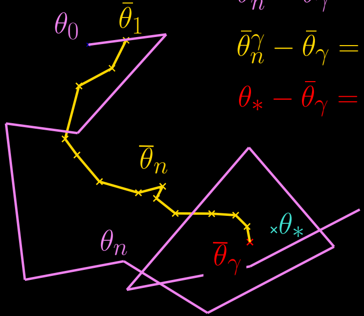
$$\theta_* - \bar\theta_\gamma = O(\gamma)$$



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

$$2\bar\theta_n^\gamma - \bar\theta_n^{2\gamma}$$

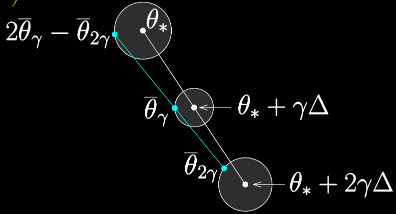
Richardson extrapolation



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

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

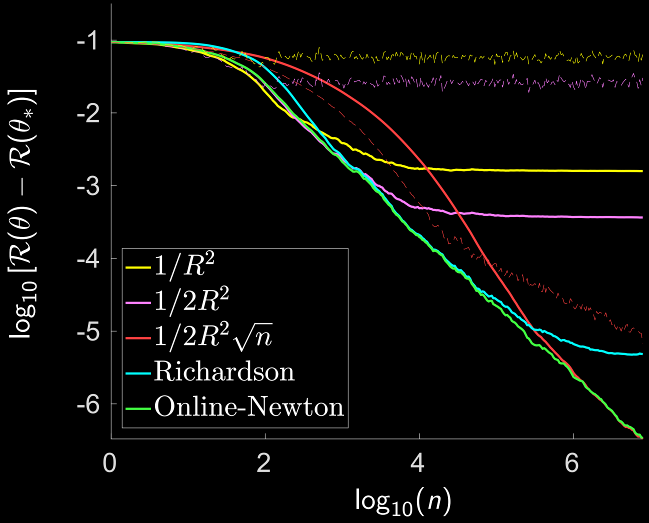
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



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

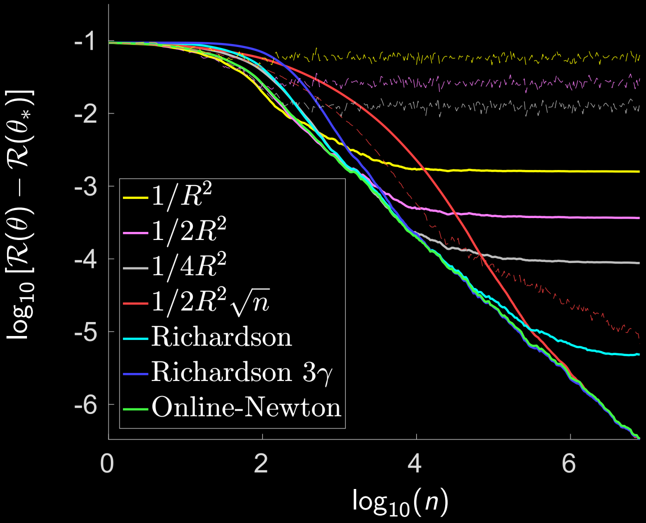
$$2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$$

Experiments: smaller dimension



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*: $\tilde{\theta}_n^3 = \frac{8}{3}\bar{\theta}_n^{\gamma} - 2\bar{\theta}_n^{2\gamma} + \frac{1}{3}\bar{\theta}_n^{4\gamma}$

Conclusion MC

Take home

- Asymptotic sometimes matter less than first iterations: consider large step size.
- Constant step size SGD is a homogeneous Markov chain.
- Difference between LS and general smooth loss is intuitive.

For smooth strongly convex loss:

- Convergence in terms of Wasserstein distance.
- Decomposition as three sources of error: variance, initial conditions, and “drift”
- Detailed analysis of the position of the limit point: the direction does not depend on γ at first order $\implies$ Extrapolation tricks can help.

Further references

Many stochastic algorithms not covered in this talk (coordinate descent, online Newton, composite optimization, non convex learning) ...

- Good introduction: Francis's lecture notes at Orsay
- Book: Convex Optimization: Algorithms and Complexity, Sébastien Bubeck

Agarwal, A., Bartlett, P. L., Ravikumar, P., and Wainwright, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249.

Agarwal, A. and Bottou, L. (2014). A Lower Bound for the Optimization of Finite Sums. *ArXiv e-prints*.

Arjevani, Y. and Shamir, O. (2016). Dimension-free iteration complexity of finite sum optimization problems. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3540–3548. 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)*.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2002). *Localized Rademacher Complexities*, pages 44–58. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014a). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654.

Defazio, A., Domke, J., and Caetano, T. (2014b). Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the 31st international conference on machine learning (ICML-14)*, pages 1125–1133.

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., Durmus, A., and Bach, F. (2017). Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *arxiv*.

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).

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Konečnỳ, J. and Richtárik, P. (2013). Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*.

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.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer.

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.

Schmidt, M., Le Roux, N., and Bach, F. (2013). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

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