Stochastic Optimization

Aymeric DIEULEVEUT

CMAP, Polytechnique

April 25, 2019

Journées Calcul et Apprentissage, Université Lyon 1

Outline

1. General context and examples.
2. What makes optimization hard?

In the context of supervised machine learning:

3. Minimizing Empirical Risk.
4. Minimizing Generalization Risk.
5. Markov chain point of view.

General context

What is optimization about?

$$\min_{\theta \in \Theta} f(\theta)$$

With θ a parameter, and f a cost function.

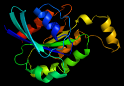
Why?
We formulate our problem as an optimization problem.
3 examples:

- Supervised machine learning
- Signal Processing
- Optimal transport

Some Examples

Example 1: Supervised Machine Learning

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



Bio-informatics

Input: DNA/RNA sequence,
Output: Drug responsiveness

Image classification

Input: Images,
Output: Digit

Supervised Machine Learning

Example 1: Supervised Machine Learning

Consider an input/output pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim \rho$.

Goal: function $\theta : \mathcal{X} \to \mathbb{R}$, s.t. $\theta(X)$ good prediction for Y.

Here, as a linear function $\langle \theta, \Phi(X) \rangle$ of features $\Phi(X) \in \mathbb{R}^d$.

Consider a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$

Define the Generalization risk :

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

Empirical Risk minimization (I)

Data: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, i.i.d.

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) : 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:

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

Some Examples

Example 2: Signal processing

Observe a signal $Y \in \mathbb{R}^{n \times q}$, try to recover the source $B \in \mathbb{R}^{p \times q}$, knowing the "forward matrix" $X \in \mathbb{R}^{n \times p}$. (multi-task regression)

$$\min_{\beta} \|X\beta - Y\|_F^2$$

Ω sparsity inducing regularization.

How to choose λ?

Some Examples

Example 3: Optimal transport

$$\min_{\pi \in \Pi} \int c(x, y) \mathrm{d}\pi(x, y)$$

Π set of probability distributions $c(x, y)$ “distance” from x to y.

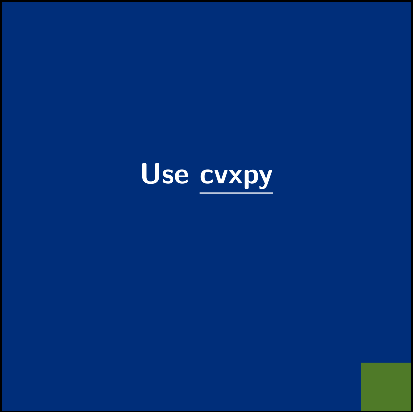
$+$ regularization

Kantorovic formulation of OT.

Is it a (hard) problem?

for convex optimization, in 99 % of the cases, no.

In other words:

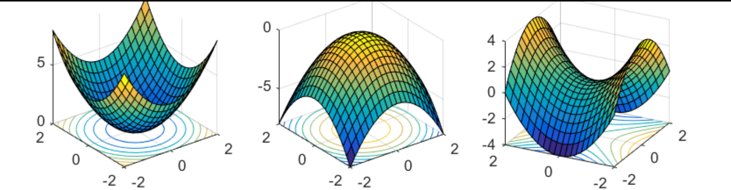


⇈

Interesting (or hard) problems

What makes it hard: 1. Convexity

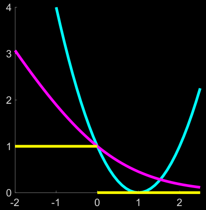
Why?



Typical non-convex problems:

Empirical risk minimization with 0-1 loss.

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



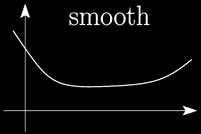
Neural networks: parametric non-convex functions.

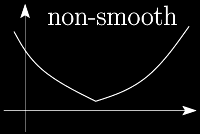
What makes it hard: 2. Regularity of the function

a. 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 $\theta \in \mathbb{R}^d$:

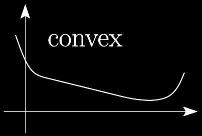
$$g(\theta) \leq g(\theta') + \langle g(\theta'), \theta - \theta' \rangle + L \left\|\theta - \theta'\right\|^2$$

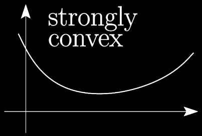
What makes it hard: 2. Regularity of the function

b. Strong Convexity

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

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





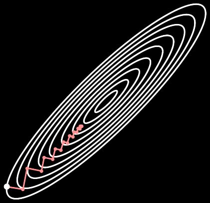
For all $\theta \in \mathbb{R}^d$:

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

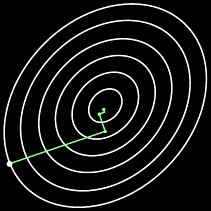
What makes it hard: 2. Regularity of the function

Why?

Rates typically depend on the condition number $\kappa = \frac{L}{\mu}$:



Large κ
harder to optimize



Small κ
easier to optimize

Smoothness and strong convexity in ML

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 $\frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^\top$

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.

Importance of regularization: provides strong convexity, and avoids overfitting.

Note: when considering dual formulation of the problem:

- L-smoothness $\leftrightarrow 1/L$-strong convexity.
- μ-strong convexity $\leftrightarrow 1/\mu$-smoothness

What makes it hard: 3. Set Θ, complexity of f

a. **Set Θ:** (if Θ is a convex set.)

- May be described implicitly (via equations):
  $\Theta = \{\theta \in \mathbb{R}^d \text{ s.t. } \|\theta\|_2 \leq R \text{ and } \langle \theta, 1 \rangle = r\}$.
  ↬ Use dual formulation of the problem.
- Projection might be difficult or impossible.
- Even when $\Theta = \mathbb{R}^d$, d might be very large (typically millions)
  ↬ use only first order methods

b. **Structure of f.** If $f = \hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \Phi(x_i) \rangle)$, computing a gradient has a cost proportional to n.

Optimization

Take home

- We express problems as minimizing a function over a set
- Most convex problems are solved
- Difficulties come from non-convexity, lack of regularity, complexity of the set Θ (or high dimension), complexity of computing gradients

What happens for supervised machine learning? Goals:

- present algorithms (convex, large dimension, high number of observations)
- show how rates depend on smoothness and strong convexity
- show how we can use the structure
- not forgetting the initial problem...!

Stochastic algorithms for ERM

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

Two fundamental questions: (a) computing (b) analyzing $\hat{\theta}$.

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

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

Gradient Descent (GD) :

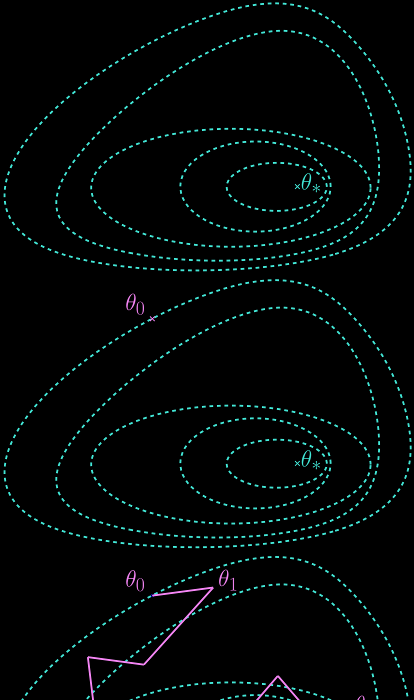
$$\boxed{\theta_k = \theta_{k-1} - \gamma_k \hat{\mathcal{R}}'(\theta_{k-1})}$$

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

2. Large $n \implies$ Stochastic algorithms

Stochastic Gradient Descent (SGD)

Stochastic Gradient des



- **Goal:**

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

  given unbiased gradient estimates f'_n
- $\theta_* := \mathrm{argmin}_{\mathbb{R}^d} f(\theta)$.

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 $I_k \sim \mathcal{U}\{1, \ldots n\}$:

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

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

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

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 $(\gamma_k)_{k\geq 0}$.

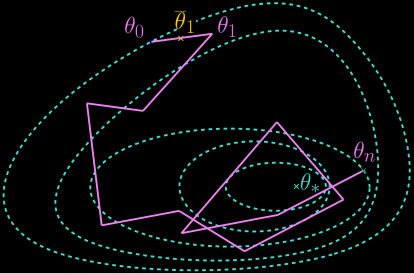
For smooth and strongly convex problem, $\theta_k \to \theta_*$ a.s. if

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

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

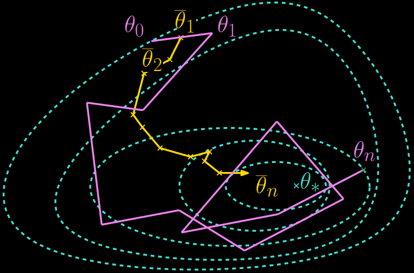
- Limit variance scales as $1/\mu^2$
- Very sensitive to ill-conditioned problems.
- μ generally unknown...

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: $\bar{\theta}_{k+1} = \frac{1}{k+1}\theta_{k+1} + \frac{k}{k+1}\bar{\theta}_k$.

Convex stochastic approximation: convergence

Known global minimax rates for non-smooth problems

- 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 $\gamma_k \propto k^{-1/2}$

For smooth problems

- Strongly convex: $O(\mu k)^{-1}$
  for $\gamma_k \propto k^{-1/2}$: adapts to strong convexity.

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

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

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

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

$\ominus$ Gradient descent update costs n times as much as SGD update.

Can we get best of both worlds?

Methods for finite sum minimization

- 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

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

↬ ⊕ update costs the same as SGD

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

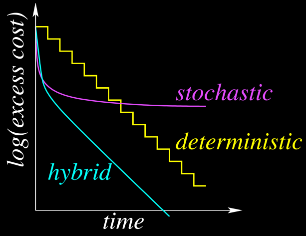
Some references:

- 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))

Take home

Stochastic algorithms for Empirical Risk Minimization.

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

What about generalization risk

Initial problem: **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.

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

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

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

	$\min \hat{\mathcal{R}}$			$\min \mathcal{R}$
	SGD	GD	SAG	SGD
Convex	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{k}\right)$		$O\left(\frac{1}{\sqrt{n}}\right)$
Stgly-Cvx	$O\left(\frac{1}{\mu k}\right)$	$O(e^{-\mu k})$	$O\left(1 - (\mu \wedge \frac{1}{n})\right)^k$	$O\left(\frac{1}{\mu n}\right)$
		$0 \leq k$		$0 \leq k \leq n$

Gradient is unknown

Least Mean Squares: rate independent of μ

Least-squares: $\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}\big[(Y - \langle \Phi(X), \theta \rangle)^2\big]$

Analysis for averaging and constant step-size $\gamma = 1/(4R^2)$ (Bach and Moulines, 2013)

- Assume $\|\Phi(x_n)\| \leqslant r$ and $|y_n - \langle \Phi(x_n), \theta_* \rangle| \leqslant \sigma$
- No assumption regarding lowest eigenvalues of the Hessian

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

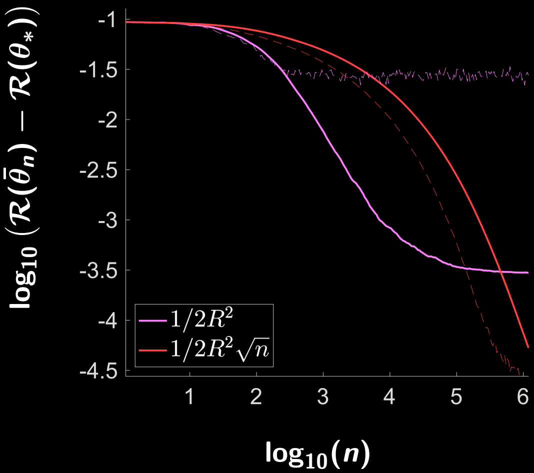
- Matches statistical lower bound (Tsybakov, 2003).
- Optimal rate with "large" step sizes

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 .

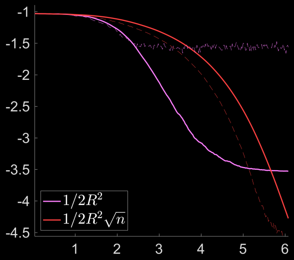
Beyond least squares. Logistic regression

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



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

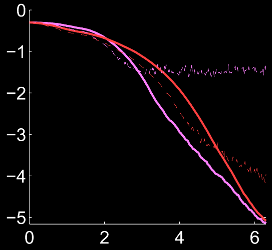
Motivation 2/ 2. Difference between quadratic and logistic loss



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:

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

- 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:

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

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

↬ Provable convergence improvement with extrapolation tricks.

[†] Dieuleveut, Durmus, Bach [2017], published in AOS 19

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 $\bar{\theta}_\gamma$?

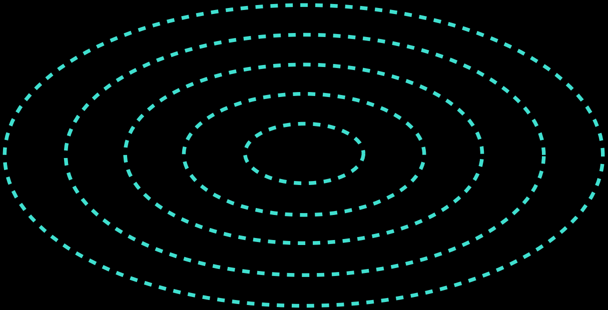
If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

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

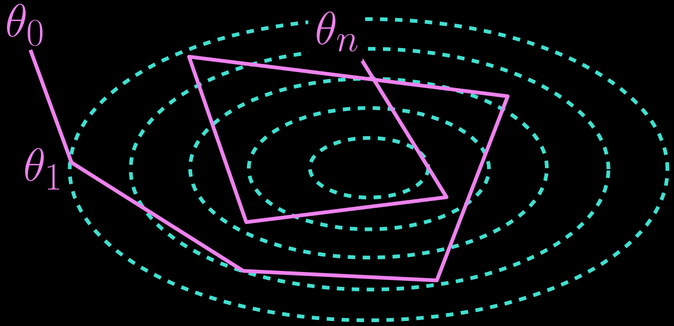
$$\mathbb{E}_{\pi_\gamma}\left[\mathcal{R}'(\theta)\right] = 0$$

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

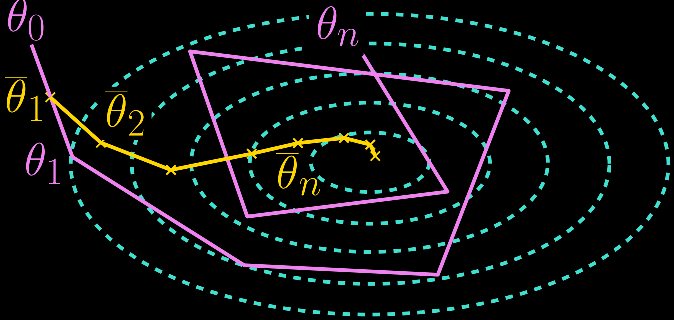
Constant learning rate SGD: convergence in the quadratic case



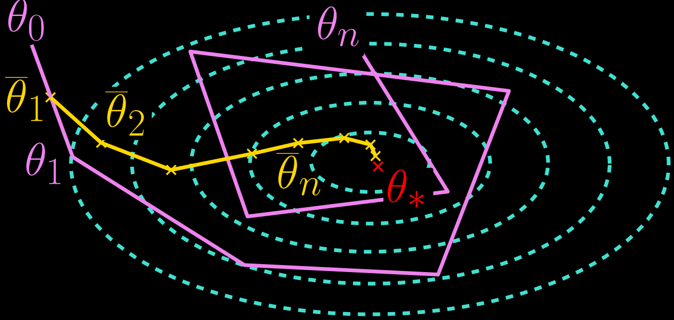
Constant learning rate SGD: convergence in the quadratic case



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 $\theta_1 \sim \pi_\gamma$.

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

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

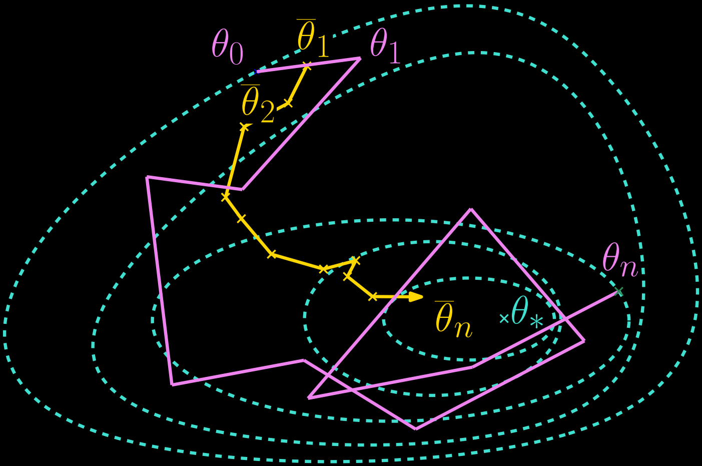
In the quadratic case (linear gradients) $\Sigma\mathbb{E}_{\pi_\gamma}[\theta - \theta_*] = 0$: $\bar{\theta}_\gamma = \theta_*$!

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

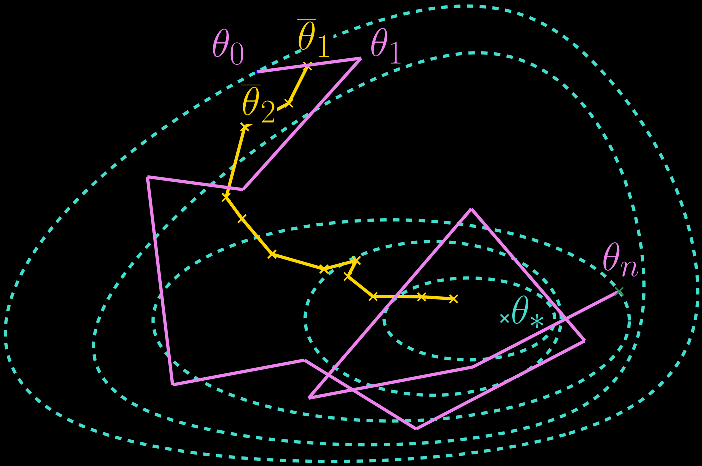
$$\bar{\theta}_\gamma - \theta_* \simeq \gamma\mathcal{R}''(\theta_*)^{-1}\mathcal{R}'''(\theta_*)\Big(\big[\mathcal{R}''(\theta_*) \otimes I + I \otimes \mathcal{R}''(\theta_*)\big]^{-1}\mathbb{E}_\varepsilon[\varepsilon(\theta_*)^{\otimes 2}]\Big)$$

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

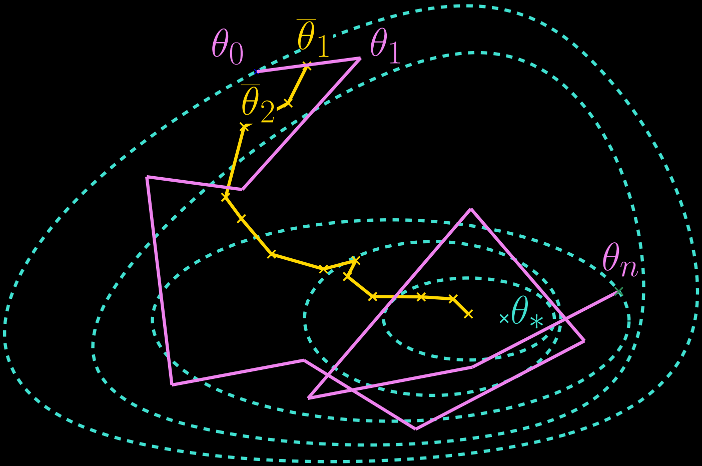
Constant learning rate SGD: convergence in the non-quadratic case



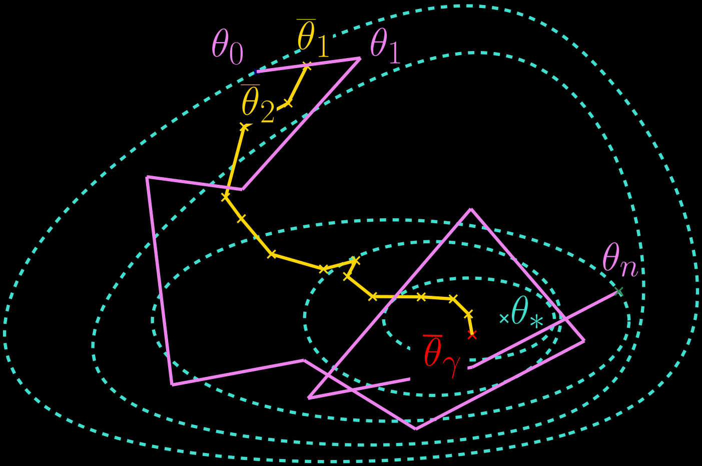
Constant learning rate SGD: convergence in the non-quadratic case



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



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

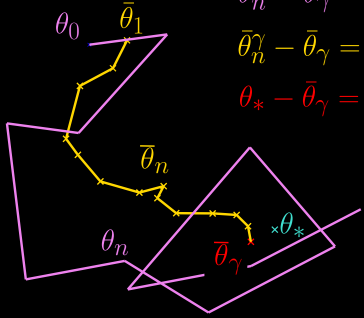


Richardson extrapolation

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

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

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

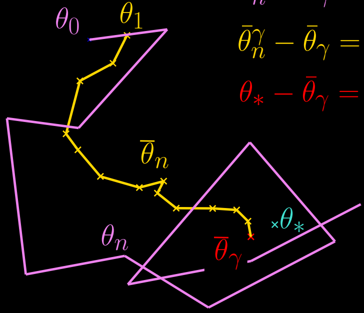


θ_*

$\longleftarrow \theta_* + \gamma\Delta$

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

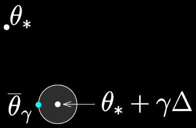
Richardson extrapolation



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

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

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



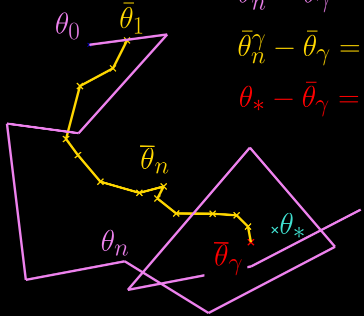
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar\theta_n^\gamma - \bar\theta_n^{2\gamma}$

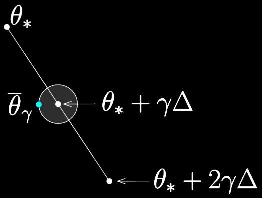
Richardson extrapolation

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

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

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





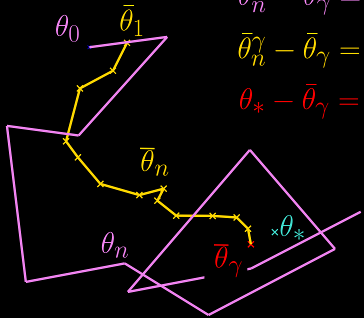
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

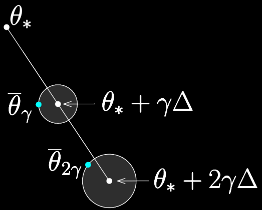
Richardson extrapolation

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

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

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





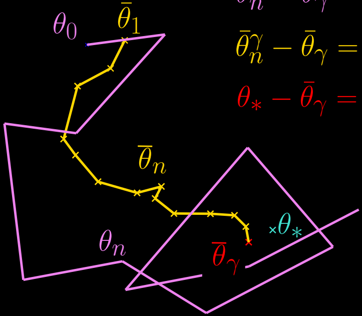
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$

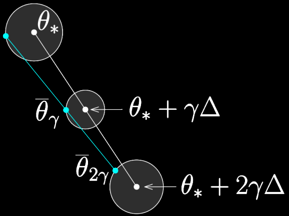
Richardson extrapolation

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

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

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





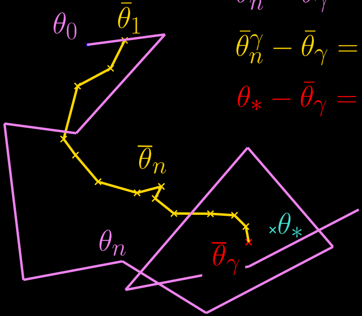
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar\theta_n^\gamma - \bar\theta_n^{2\gamma}$

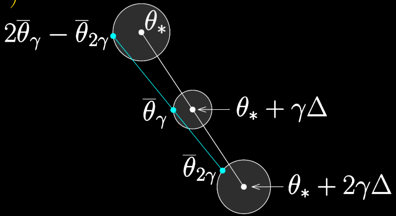
Richardson extrapolation

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

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

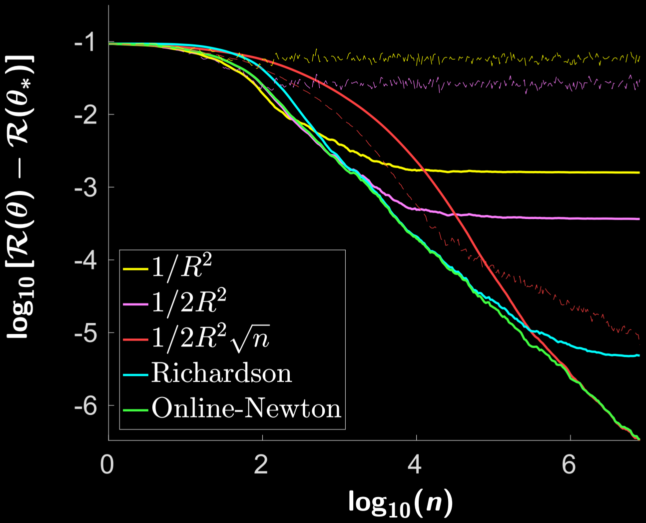
$$\theta_* - \bar\theta_\gamma = O(\gamma)$$





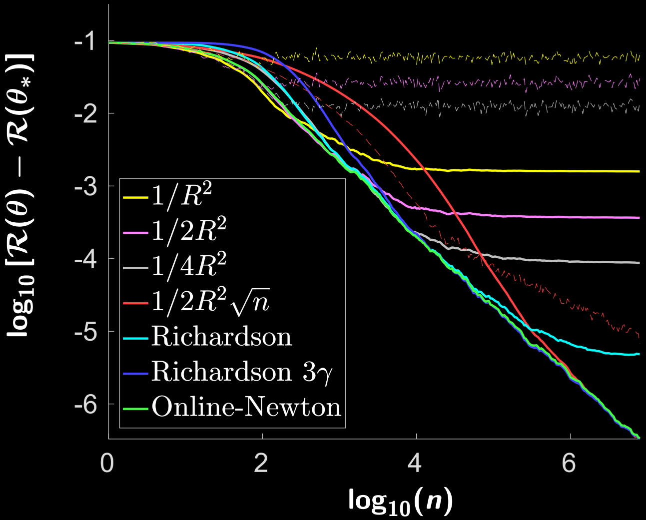
Recovering convergence closer to θ_* by Richardson extrapolation $2\bar\theta_n^\gamma - \bar\theta_n^{2\gamma}$

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 $\Longrightarrow$ 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

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.

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.

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.

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.