Stochastic Approximation in Hilbert Spaces

Aymeric DIEULEVEUT

Supervised by Francis BACH

September 28, 2017

Outline

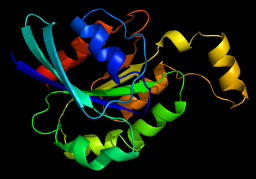
1. Introduction:
   - Supervised Machine Learning
   - Stochastic Approximation
2. Finite dimensional results
3. Infinite dimensional results
4. Beyond quadratic loss: interpretation as a Markov chain.

Supervised Machine Learning: definition & applications

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

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

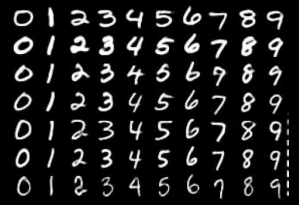
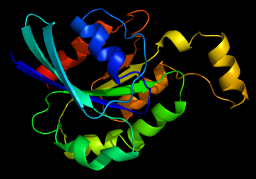


Image classification

Input: Handwritten digits / Images,
Output: Digit

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

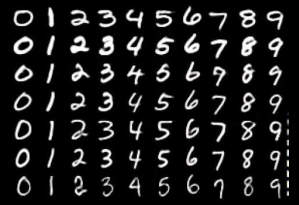


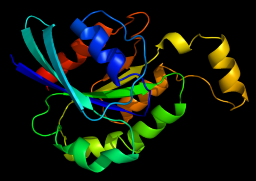
Image classification

Input: Handwritten digits / Images,
Output: Digit

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

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) $\to 10^6$

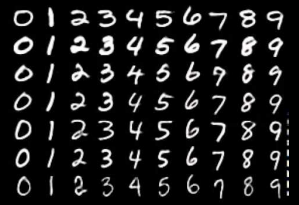


Image classification

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: mathematical framework

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

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

Supervised Machine Learning: mathematical framework

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

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

Goal: find $g : \mathcal{X} \to \mathbb{R}$, such that $g(X)$ is a *good* prediction for Y.

Supervised Machine Learning: mathematical framework

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

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

Goal: find $g : \mathcal{X} \to \mathbb{R}$, such that $g(X)$ is a *good* prediction for Y.

Measure accuracy with a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$: squared loss, logistic loss...

Supervised Machine Learning: mathematical framework

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

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

Goal: find $g : \mathcal{X} \to \mathbb{R}$, such that $g(X)$ is a *good* prediction for Y.

Measure accuracy with a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$: squared loss, logistic loss...

Risk (generalization error):

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

Parametric case: Prediction as a **linear function** $g_\theta(X) = \langle \theta, \Phi(X) \rangle$ of features $\Phi(X) \in \mathbb{R}^d$. Notation: $\mathcal{R}(\theta) := \mathcal{R}(g_\theta)$.

Supervised Machine Learning: mathematical framework

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

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

Goal: find $g : \mathcal{X} \to \mathbb{R}$, such that $g(X)$ is a *good* prediction for Y.

Measure accuracy with a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$: squared loss, logistic loss...

Risk (generalization error):

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

Parametric case: Prediction as a **linear function** $g_\theta(X) = \langle \theta, \Phi(X) \rangle$ of features $\Phi(X) \in \mathbb{R}^d$. Notation: $\mathcal{R}(\theta) := \mathcal{R}(g_\theta)$.
Non-parametric case: Prediction as a function $g \in \mathcal{H}$, for $\mathcal{H}$ infinite-dimensional space.

Empirical Risk minimization (I) - Parametric case

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, **i.i.d.**
- Empirical risk (or training error):

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

- First approach: **Empirical risk minimization (regularized)**:

$$\hat{\theta} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \quad \hat{\mathcal{R}}(\theta) \quad + \quad \mu\Omega(\theta).$$

data fitting term + regularizer

Empirical Risk minimization (II) - Parametric case

- For example, least-squares regression:

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

Empirical Risk minimization (II) - Parametric case

- For example, least-squares regression:

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

- and logistic regression:

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

Empirical Risk minimization (II) - Parametric case

- For example, least-squares regression:

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

- and logistic regression:

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

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

Empirical Risk minimization (II) - Parametric case

- For example, least-squares regression:

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

- and logistic regression:

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

- **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. True risk is more important than empirical risk.

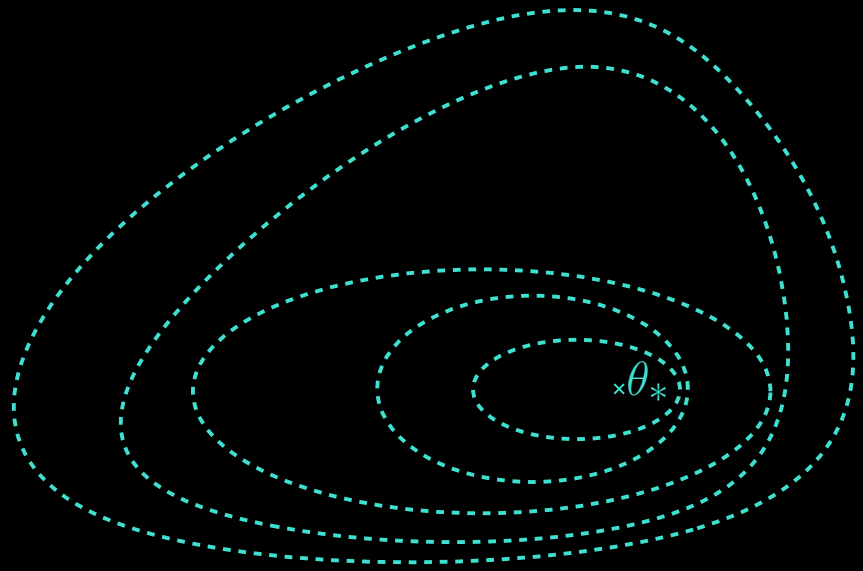
Stochastic Approximation

- **Goal**:

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

  given unbiased gradient estimates f'_n

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



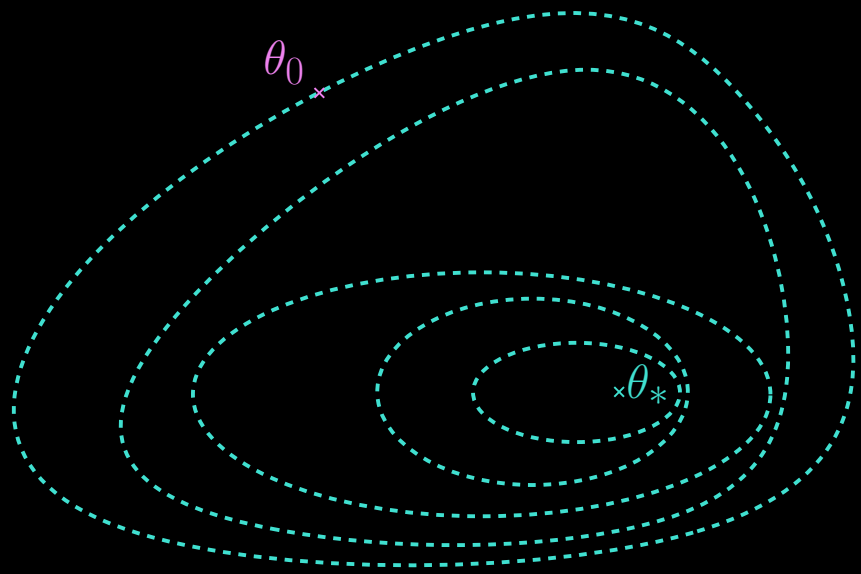
Stochastic Approximation

- **Goal**:

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

given unbiased gradient estimates f'_n

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



- **Key algorithm: Stochastic Gradient Descent (SGD)** [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.

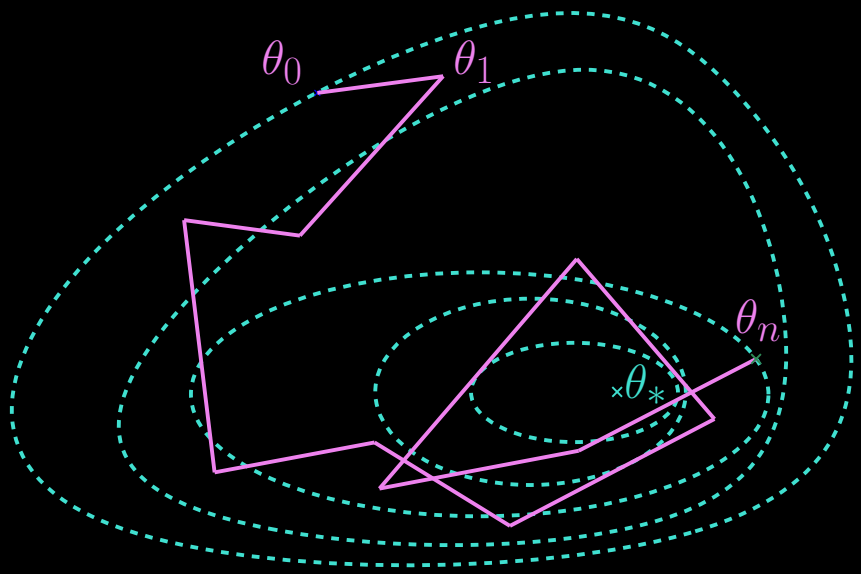
Stochastic Approximation

- **Goal**:

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

given unbiased gradient estimates f'_n

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



- **Key algorithm: Stochastic Gradient Descent (SGD)** [Robbins and Monro, 1951]:

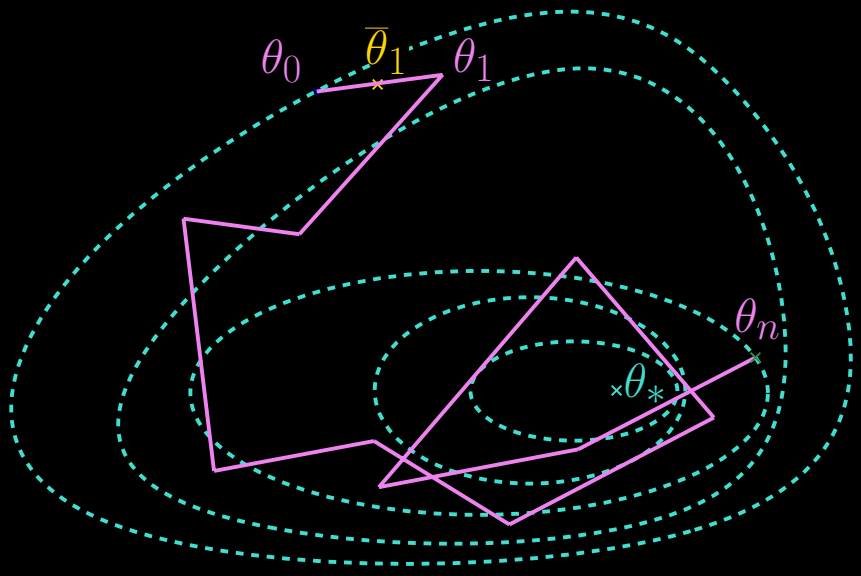
$$\boxed{\theta_n = \theta_{n-1} - \gamma_n f'_n(\theta_{n-1})}$$

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

Polyak Ruppert averaging

Introduced by Polyak and Juditsky [1992] and Ruppert [1988]:

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

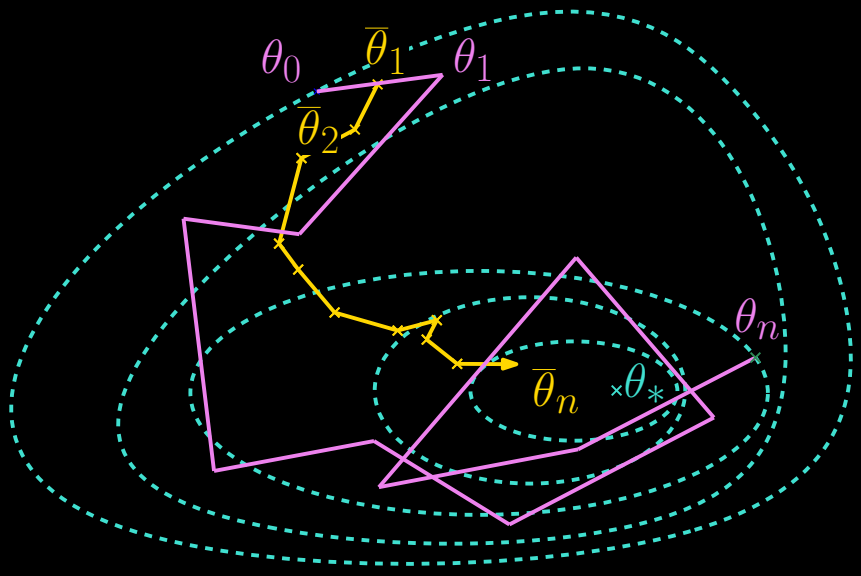


- off line averaging reduces the noise effect.

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.

Stochastic Approximation (SA) in Machine Learning

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

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

SA for the **true risk** :

- 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}:

$$\begin{aligned} \mathcal{R}(\theta) &= \mathbb{E}\, \ell(y_k, \langle \theta, \Phi(x_k) \rangle) \\ f_k'(\theta_{k-1}) &= \ell'(y_k, \langle \theta_{k-1}, \Phi(x_k) \rangle) \end{aligned}$$

Stochastic Approximation (SA) in Machine Learning

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

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

SA for the **true risk** :

- 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}:

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

Stochastic Approximation (SA) in Machine Learning

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

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

SA for the **true risk** :

- 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}:

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

Single pass through the data – "Automatic" regularization.

Central algorithm in the thesis.

Outline: bibliography

a) *Non-parametric Stochastic Approximation with Large Step-sizes,*
A. Dieuleveut and F. Bach, in the Annals of Statistics

b) *Harder, Better, Faster, Stronger Convergence Rates for Least-squares Regression,*
A. Dieuleveut, N. Flammarion and F. Bach, in Journal of Machine Learning Research

c) *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains,*
A. Dieuleveut, A. Durmus, F. Bach, under submission.

Outline: bibliography

a) *Non-parametric Stochastic Approximation with Large Step-sizes,*
A. Dieuleveut and F. Bach, in the Annals of Statistics

b) *Harder, Better, Faster, Stronger Convergence Rates for Least-squares Regression,*
A. Dieuleveut, N. Flammarion and F. Bach, in Journal of Machine Learning Research

c) *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains,*
A. Dieuleveut, A. Durmus, F. Bach, under submission.

	Quadratic loss	Smooth loss	FD	Non-parametric
a)	✓		✓	✓
b)	✓		✓	✓
c)	✓	✓	✓	

Outline: bibliography

a) *Non-parametric Stochastic Approximation with Large Step-sizes,*
A. Dieuleveut and F. Bach, in the Annals of Statistics

b) *Harder, Better, Faster, Stronger Convergence Rates for Least-squares Regression,*
A. Dieuleveut, N. Flammarion and F. Bach, in Journal of Machine Learning Research

c) *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains,*
A. Dieuleveut, A. Durmus, F. Bach, under submission.

	Quadratic loss	Smooth loss	FD	Non-parametric
a)	✓		✓	✓
b)	✓		✓	✓
c)	✓	✓	✓	

Part 1

Outline: bibliography

a) *Non-parametric Stochastic Approximation with Large Step-sizes,*
A. Dieuleveut and F. Bach, in the Annals of Statistics

b) *Harder, Better, Faster, Stronger Convergence Rates for Least-squares Regression,*
A. Dieuleveut, N. Flammarion and F. Bach, in Journal of Machine Learning Research

c) *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains,*
A. Dieuleveut, A. Durmus, F. Bach, under submission.

	Quadratic loss	Smooth loss	FD	Non-parametric
a)	✓		✓	✓
b)	✓		✓	✓
c)	✓	✓	✓	

Part 1 – Part 2

Outline: bibliography

a) *Non-parametric Stochastic Approximation with Large Step-sizes,*
A. Dieuleveut and F. Bach, in the Annals of Statistics

b) *Harder, Better, Faster, Stronger Convergence Rates for Least-squares Regression,*
A. Dieuleveut, N. Flammarion and F. Bach, in Journal of Machine Learning Research

c) *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains,*
A. Dieuleveut, A. Durmus, F. Bach, under submission.

	Quadratic loss	Smooth loss	FD	Non-parametric
a)	✓		✓	✓
b)	✓		✓	✓
c)	✓	✓	✓	

Part 1 – Part 2 – Part 3

Outline

1. Introduction.
2. A warm up! Results in finite dimension, ($d \gg n$)
   - Averaged stochastic descent: adaptivity
   - Acceleration: two optimal rates
3. Non-parametric stochastic approximation
4. Stochastic approximation as a Markov chain: extension to non quadratic loss functions.

Behavior of Stochastic Approximation in high dimension

Least-squares regression in finite dimension:

$$\mathcal{R}(\theta) = \mathbb{E}_\rho \left[\Big(\langle \theta, \Phi(X) \rangle - Y \Big)^2 \right].$$

Behavior of Stochastic Approximation in high dimension

Least-squares regression in finite dimension:

$$\mathcal{R}(\theta) = \mathbb{E}_\rho \left[\Big(\langle \theta, \Phi(X) \rangle - Y \Big)^2 \right].$$

Let $\Sigma = \mathbb{E}\left[\Phi(X)\Phi(X)^\top\right] \in \mathbb{R}^{d \times d}$: for θ_* the best linear predictor,

$$\mathcal{R}(\theta) - \mathcal{R}(\theta_*) = \left\| \Sigma^{1/2}(\theta - \theta_*) \right\|^2.$$

Let $R^2 := \mathbb{E}\left[\|\Phi(X)\|^2\right]$, $\sigma^2 := \mathbb{E}\left[(Y - \langle \theta_*, \Phi(X) \rangle)^2\right]$.

Behavior of Stochastic Approximation in high dimension

Least-squares regression in finite dimension:

$$\mathcal{R}(\theta) = \mathbb{E}_\rho \left[\left(\langle \theta, \Phi(X) \rangle - Y \right)^2 \right].$$

Let $\Sigma = \mathbb{E}\left[\Phi(X)\Phi(X)^\top\right] \in \mathbb{R}^{d \times d}$: for θ_* the best linear predictor,

$$\mathcal{R}(\theta) - \mathcal{R}(\theta_*) = \left\| \Sigma^{1/2}(\theta - \theta_*) \right\|^2.$$

Let $R^2 := \mathbb{E}\left[\|\Phi(X)\|^2\right]$, $\sigma^2 := \mathbb{E}\left[(Y - \langle \theta_*, \Phi(X) \rangle)^2\right]$.

Consider stochastic gradient descent (*a.k.a., Least-Mean-Squares*)

Theorem

For any $\gamma \leq \frac{1}{4R^2}$, for any $\alpha > 1$, for any $r \geq 0$, for any $n \in \mathbb{N}$,

$$\mathbb{E}\mathcal{R}\left(\bar{\theta}_n\right) - \mathcal{R}(\theta_*) \leq \frac{4\sigma^2 \gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}} + \frac{4 \left\| \Sigma^{1/2-r}(\theta_* - \theta_0) \right\|^2}{\gamma^{2r} n^{\min(2r,2)}}.$$

Theorem 1[†], consequences

Theorem

For any $\gamma \leq \frac{1}{4R^2}$, for any $\alpha > 1$, for any $r \geq 0$, for any $n \in \mathbb{N}$,

$$\mathbb{E}\mathcal{R}\left(\bar{\theta}_n\right) - \mathcal{R}(\theta_*) \leq \underbrace{\frac{4\sigma^2\gamma^{1/\alpha}\operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}}}_{\text{Variance}} + \underbrace{\frac{4\left\|\Sigma^{1/2-r}(\theta_* - \theta_0)\right\|^2}{\gamma^{2r}n^{\min(2r,2)}}}_{\text{Bias}}.$$

[†] Dieuleveut and Bach [2015].

Theorem 1[†], consequences

Theorem
For any $\gamma \leq \frac{1}{4R^2}$, for any $\alpha > 1$, for any $r \geq 0$, for any $n \in \mathbb{N}$,

$$\mathbb{E}\mathcal{R}\left(\bar{\theta}_n\right) - \mathcal{R}(\theta_*) \leq \underbrace{\frac{4\sigma^2\gamma^{1/\alpha}\operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}}}_{\text{Variance}} + \underbrace{\frac{4\left\|\Sigma^{1/2-r}(\theta_* - \theta_0)\right\|^2}{\gamma^{2r}n^{\min(2r,2)}}}_{\text{Bias}}.$$

Variance term		Bias term	
$\gamma\sigma^2\operatorname{tr}(\Sigma)$	$\frac{\sigma^2 d}{n}$	$\frac{\|\theta_* - \theta_0\|^2}{\gamma n}$	$\frac{\left\|\Sigma^{-1/2}(\theta_* - \theta_0)\right\|^2}{\gamma^2 n^2}$
$\alpha = 1$	$\alpha \to \infty$	$r = 1/2.$	$r = 1.$

[†] Dieuleveut and Bach [2015].

Theorem 1[†], consequences

Theorem
For any $\gamma \leq \frac{1}{4R^2}$, for any $\alpha > 1$, for any $r \geq 0$, for any $n \in \mathbb{N}$,

$$\mathbb{E}\mathcal{R}\left(\bar{\theta}_n\right) - \mathcal{R}(\theta_*) \leq \underbrace{\frac{4\sigma^2 \gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}}}_{\text{Variance}} + \underbrace{\frac{4 \left\|\Sigma^{1/2-r}(\theta_* - \theta_0)\right\|^2}{\gamma^{2r} n^{\min(2r,2)}}}_{\text{Bias}}.$$

Variance term		Bias term	
$\gamma\sigma^2 \operatorname{tr}(\Sigma)$	$\frac{\sigma^2 d}{n}$	$\frac{\|\theta_* - \theta_0\|^2}{\gamma n}$	$\frac{\left\|\Sigma^{-1/2}(\theta_* - \theta_0)\right\|^2}{\gamma^2 n^2}$
$\alpha = 1$	$\alpha \to \infty$	$r = 1/2.$	$r = 1.$
	Recovers Bach and Moulines [2013]		Improves asymptotic Bias

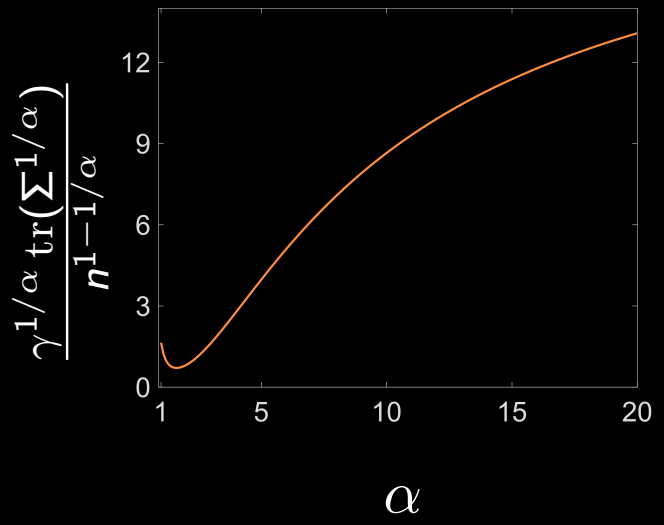
[†] Dieuleveut and Bach [2015].

Theorem 1, consequences

Theorem

For any $\gamma \leq \frac{1}{4R^2}$, for any $n \in \mathbb{N}$,

$$\mathbb{E}\mathcal{R}\left(\bar{\theta}_n\right) - \mathcal{R}(\theta_*) \leq \inf_{\alpha>1, r\geq 0} \left(\underbrace{\frac{4\sigma^2 \gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}}}_{\text{Variance}} + \underbrace{\frac{4 \left\| \Sigma^{1/2-r}(\theta_* - \theta_0) \right\|^2}{\gamma^{2r} n^{\min(2r,2)}}}_{\text{Bias}} \right).$$



Adaptivity

Upper bound on the variance term as a function of α.

$d \gg n$.

Limits to SA performance: two lower bounds

Stochastic Approximation in Supervised ML

Limits to SA performance: two lower bounds

Stochastic Approximation in Supervised ML

Builds an estimator **given** n **observations.**

$\looparrowright$ statistical **lower bound**:

$$\frac{\sigma^2 d}{n}$$

Limits to SA performance: two lower bounds

Stochastic Approximation in Supervised ML

Builds an estimator **given** n **observations.**
$\hookrightarrow$ statistical **lower bound**:

$$\frac{\sigma^2 d}{n}$$

Approximates the minimum of an ($L-$smooth) function **in** t **iterations**, using first order information.
$\hookrightarrow$ optimization **lower bound**:

$$\frac{L \left\| \theta_0 - \theta_* \right\|^2}{t^2}.$$

Limits to SA performance: two lower bounds

Stochastic Approximation in Supervised ML

Builds an estimator **given** n **observations.**

$\hookrightarrow$ statistical **lower bound**:

$$\frac{\sigma^2 d}{n}$$

Approximates the minimum of an ($L-$smooth) function **in** t **iterations**, using first order information.

$\hookrightarrow$ optimization **lower bound**:

$$\frac{L \left\| \theta_0 - \theta_* \right\|^2}{t^2}.$$

here, $n = t$.

Limits to SA performance: two lower bounds

Stochastic Approximation in Supervised ML

Builds an estimator **given** n **observations.**

$\hookrightarrow$ statistical **lower bound**:

$$\frac{\sigma^2 d}{n}$$

Approximates the minimum of an ($L-$smooth) function **in** n **iterations**, using first order information.

$\hookrightarrow$ optimization **lower bound**:

$$\frac{L \left\| \theta_0 - \theta_* \right\|^2}{n^2}.$$

here, $n = t$.

Limits to SA performance: two lower bounds

Stochastic Approximation in Supervised ML

Builds an estimator **given** n **observations.**
$\hookrightarrow$ statistical **lower bound**:

$$\frac{\sigma^2 d}{n}$$

Approximates the minimum of an ($L-$smooth) function **in** n **iterations**, using first order information.
$\hookrightarrow$ optimization **lower bound**:

$$\frac{L \left\| \theta_0 - \theta_* \right\|^2}{n^2}.$$

here, $n = t$.

Theorem 1, for Av-SGD, gives as upper bound:

$$\frac{\sigma^2 d}{n} + \min \left(\frac{L \left\| \theta_0 - \theta_* \right\|^2}{n}; \frac{L^2 \left\| \Sigma^{-1/2} (\theta_0 - \theta_*) \right\|^2}{n^2} \right) .$$

Acceleration[†]

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

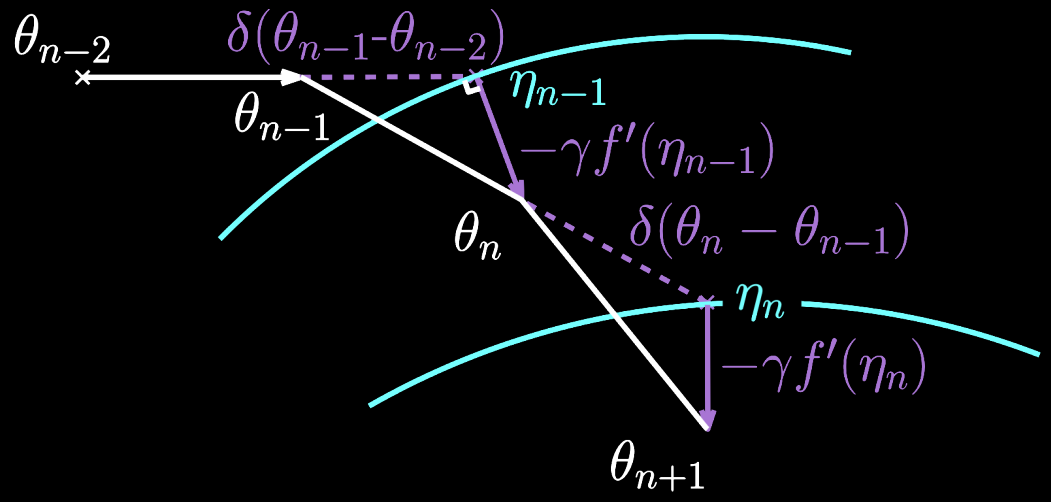
$$\left\{ \begin{array}{rcl} \theta_n & = & \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n & = & \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{array} \right.$$

[†] *Dieuleveut, Flammarion, Bach [2016]*

Acceleration[†]

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{cases}$$

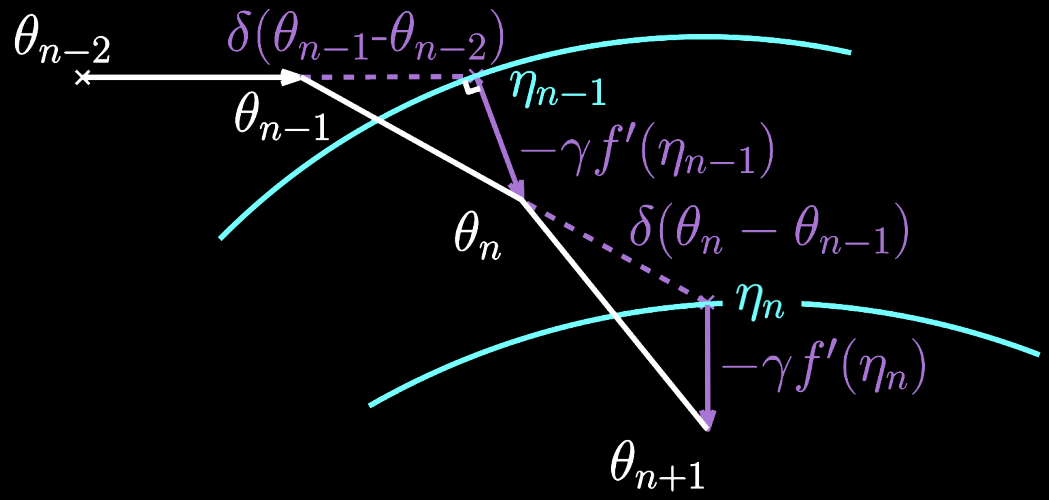


[†] *Dieuleveut, Flammarion, Bach [2016]*

Acceleration[†]

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{cases}$$



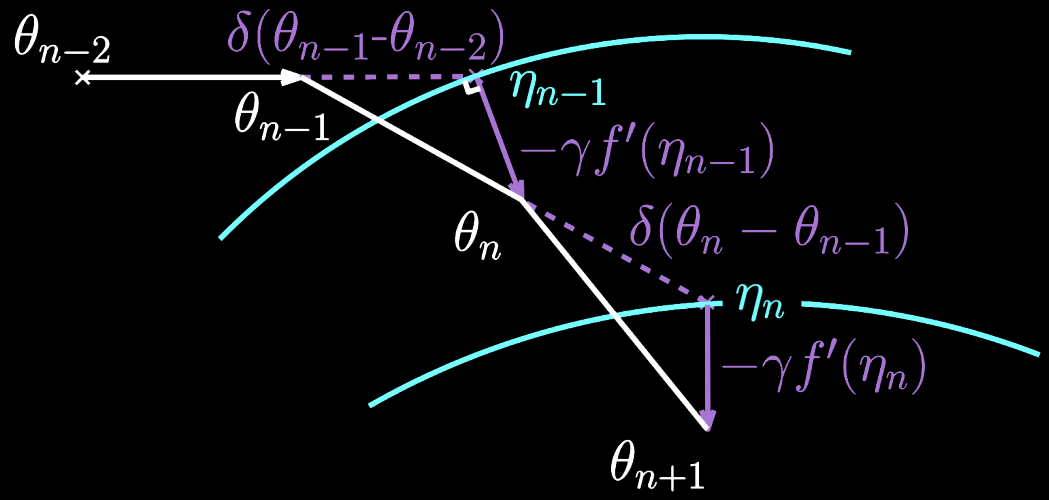
Problem: acceleration is sensitive to noise [d'Aspremont, 2008].

[†] *Dieuleveut, Flammarion, Bach [2016]*

Acceleration†

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{cases}$$



Problem: acceleration is sensitive to noise [d'Aspremont, 2008].

Combining SGD, acceleration and averaging,

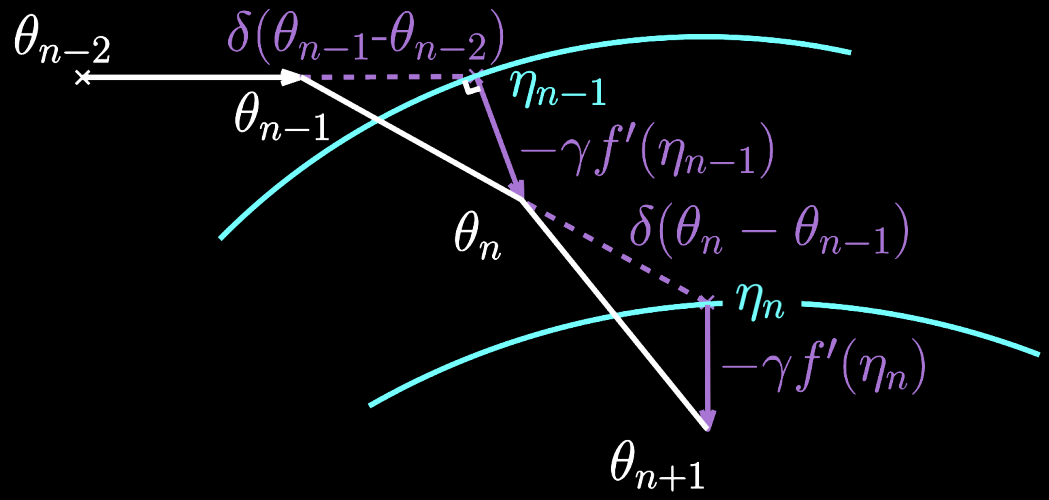
- using extra regularization,

† *Dieuleveut, Flammarion, Bach [2016]*

Acceleration†

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{cases}$$



Problem: acceleration is sensitive to noise [d'Aspremont, 2008].

Combining SGD, acceleration and averaging,

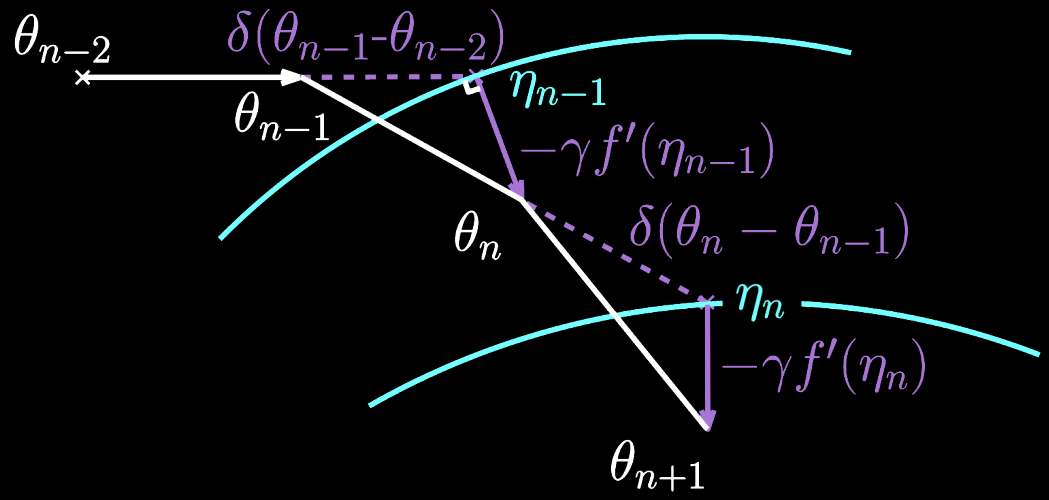
- using extra regularization,
- and for "additive" noise model only,

† *Dieuleveut, Flammarion, Bach [2016]*

Acceleration†

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{cases}$$



Problem: acceleration is sensitive to noise [d'Aspremont, 2008].

Combining SGD, acceleration and averaging,

- using extra regularization,
- and for "additive" noise model only,

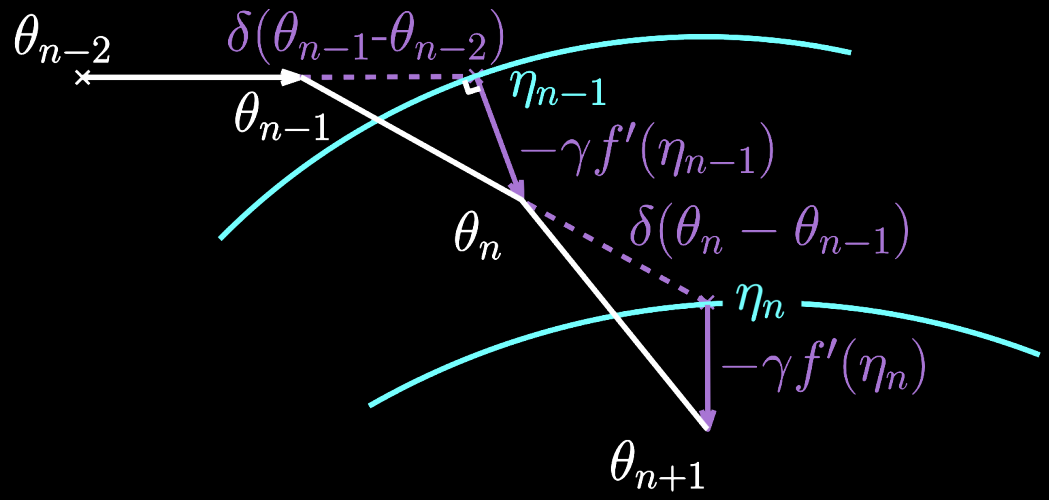
Caveat: LMS recursion does not provide an additive noise oracle. Different recursion with Σ known.

† *Dieuleveut, Flammarion, Bach [2016]*

Acceleration†

Optimal rate *(for deterministic optimization)*, is achieved by **accelerated gradient descent**:

$$\begin{cases} \theta_n &= \eta_{n-1} - \gamma_n f'(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}) \, . \end{cases}$$



Problem: acceleration is sensitive to noise [d'Aspremont, 2008].

Combining SGD, acceleration and averaging,

- using extra regularization,
- and for "additive" noise model only,

we achieve both of the optimal rates.

Caveat: LMS recursion does not provide an additive noise oracle.

Different recursion with Σ known.

† *Dieuleveut, Flammarion, Bach [2016]*

Acceleration and averaging

More precisely we consider:

$$\begin{aligned}\theta_n &= \nu_{n-1} - \gamma \mathcal{R}_n'(\nu_{n-1}) - \gamma\lambda(\nu_{n-1} - \theta_0) \\ \nu_n &= \theta_n + \delta\big(\theta_n - \theta_{n-1}\big),\end{aligned}$$

Theorem
For any $\gamma \leq 1/2R^2$, for $\delta = 1$, and $\lambda = 0$,

$$\mathbb{E}\left[\mathcal{R}(\bar{\theta}_n)\right] - \mathcal{R}(\theta_*) \leq 8\frac{\sigma^2 d}{n+1} + 36\frac{\|\theta_0 - \theta_*\|^2}{\gamma(n+1)^2}.$$

Optimal rate from both statistical and optimization point of view.

Outline

1. Introduction.
2. A warm up! Results in finite dimension, ($d \gg n$)
3. Non-parametric stochastic approximation
   - Averaged stochastic descent: statistical rate of convergence
   - Acceleration: improving convergence in ill-conditioned regimes
4. Stochastic approximation as a Markov chain: extension to non quadratic loss functions.

Non-parametric Random Design Least Squares Regression

Goal:

$$\min_g \mathcal{R}(g) = \mathbb{E}_\rho \left[(Y - g(X))^2 \right]$$

Non-parametric Random Design Least Squares Regression

Goal:

$$\min_g \mathcal{R}(g) = \mathbb{E}_\rho \left[(Y - g(X))^2 \right]$$

- ρ_X marginal distribution of X in $\mathcal{X}$,
- $L^2_{\rho_X}$ set of squared integrable functions w.r.t. ρ_X.

Non-parametric Random Design Least Squares Regression

Goal:

$$\min_g \mathcal{R}(g) = \mathbb{E}_\rho \left[(Y - g(X))^2 \right]$$

- ρ_X marginal distribution of X in $\mathcal{X}$,
- $L^2_{\rho_X}$ set of squared integrable functions w.r.t. ρ_X.

Bayes predictor minimizes the quadratic risk over $L^2_{\rho_X}$:

$$g_\rho(X) = \mathbb{E}\left[Y | X \right].$$

Non-parametric Random Design Least Squares Regression

Goal:

$$\min_g \mathcal{R}(g) = \mathbb{E}_\rho \left[(Y - g(X))^2 \right]$$

- ρ_X marginal distribution of X in $\mathcal{X}$,
- $L^2_{\rho_X}$ set of squared integrable functions w.r.t. ρ_X.

Bayes predictor minimizes the quadratic risk over $L^2_{\rho_X}$:

$$g_\rho(X) = \mathbb{E}\left[Y | X \right].$$

Moreover, for any function g in $L^2_{\rho_X}$, the excess risk is:

$$\mathcal{R}(g) - \mathcal{R}(g_\rho) = \|g - g_\rho\|^2_{L^2_{\rho_X}} \quad .$$

Non-parametric Random Design Least Squares Regression

Goal:

$$\min_g \mathcal{R}(g) = \mathbb{E}_\rho \left[(Y - g(X))^2 \right]$$

- ρ_X marginal distribution of X in $\mathcal{X}$,
- $L^2_{\rho_X}$ set of squared integrable functions w.r.t. ρ_X.

Bayes predictor minimizes the quadratic risk over $L^2_{\rho_X}$:

$$g_\rho(X) = \mathbb{E}\left[Y | X \right].$$

Moreover, for any function g in $L^2_{\rho_X}$, the excess risk is:

$$\mathcal{R}(g) - \mathcal{R}(g_\rho) = \|g - g_\rho\|^2_{L^2_{\rho_X}} .$$

$\mathcal{H}$ a space of functions: there exists $g_{\mathcal{H}} \in \bar{\mathcal{H}}^{L^2_{\rho_X}}$ such that

$$\mathcal{R}(g_{\mathcal{H}}) = \inf_{g \in \mathcal{H}} \mathcal{R}(g).$$

Reproducing Kernel Hilbert Space

Definition

A Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ is a space of functions from $\mathcal{X}$ into $\mathbb{R}$, such that there exists a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, satisfying:

- For any $x \in \mathcal{X}$, $\mathcal{H}$ contains the function K_x, defined by:

$$K_x : \mathcal{X} \to \mathbb{R}$$
$$z \mapsto K(x, z).$$

Reproducing Kernel Hilbert Space

Definition

A Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ is a space of functions from $\mathcal{X}$ into $\mathbb{R}$, such that there exists a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, satisfying:

- For any $x \in \mathcal{X}$, $\mathcal{H}$ contains the function K_x, defined by:

$$K_x : \mathcal{X} \to \mathbb{R}$$
$$z \mapsto K(x, z).$$

- For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the *reproducing property* holds:

$$\langle K_x, f \rangle_{\mathcal{H}} = f(x).$$

Why are RKHS so nice?

- **Computation:**
  - Linear spaces of functions.
  - Existence of gradients (Hilbert).
  - Possible to compute inner products thanks to the reproducing property.
  - Only deal with functions in the set $\text{span}\{K_{x_i}, i = 1 \dots n\}$ (representer theorem).

  $\looparrowright$ the algebraic framework is preserved !

Why are RKHS so nice?

- **Computation:**
  - Linear spaces of functions.
  - Existence of gradients (Hilbert).
  - Possible to compute inner products thanks to the reproducing property.
  - Only deal with functions in the set $\text{span}\{K_{x_i}, i = 1 \dots n\}$ (representer theorem).

  $\leadsto$ the algebraic framework is preserved !
- **Approximation:** many kernels satisfy $\bar{\mathcal{H}}^{L^2_{\rho_X}} = L^2_{\rho_X}$, there is no approximation error !

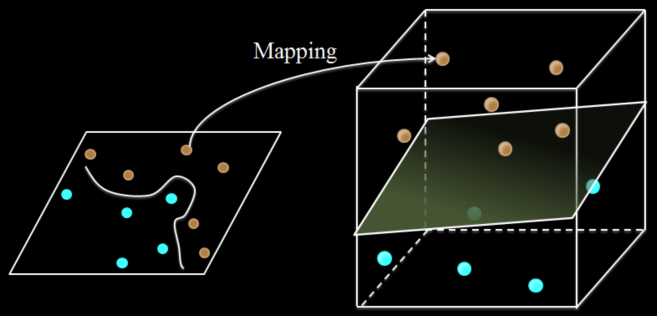
Why are RKHS so nice?

- **Computation:**
  - Linear spaces of functions.
  - Existence of gradients (Hilbert).
  - Possible to compute inner products thanks to the reproducing property.
  - Only deal with functions in the set $\text{span}\{K_{x_i}, i = 1 \dots n\}$ (representer theorem).

  $\leadsto$ the algebraic framework is preserved !
- **Approximation:** many kernels satisfy $\bar{\mathcal{H}}^{L^2_{\rho_X}} = L^2_{\rho_X}$, there is no approximation error !
- **Representation:** Feature map,

$$\begin{array}{ccc} \mathcal{X} & \to & \mathcal{H} \\ x & \mapsto & K_x \end{array}$$

maps points from *any* set into a linear space to apply a linear method.



Stochastic approximation in the RKHS.

As $\mathcal{R}(g) = \mathbb{E}\left[(\langle g, K_X\rangle_{\mathcal{H}} - Y)^2\right]$, for each pair of observations

$$(\langle g, K_{x_n}\rangle_{\mathcal{H}} - y_n)K_{x_n} = (g(x_n) - y_n)K_{x_n}$$

is an *unbiased stochastic gradient* of $\mathcal{R}$ at g.

Stochastic approximation in the RKHS.

As $\mathcal{R}(g) = \mathbb{E}\left[(\langle g, K_X\rangle_{\mathcal{H}} - Y)^2\right]$, for each pair of observations

$$(\langle g, K_{x_n}\rangle_{\mathcal{H}} - y_n)K_{x_n} = (g(x_n) - y_n)K_{x_n}$$

is an *unbiased stochastic gradient* of $\mathcal{R}$ at g.

Consider the stochastic gradient recursion, starting from $g_0 \in \mathcal{H}$:

$$g_n = g_{n-1} - \gamma\left[\langle g_{n-1}, K_{x_n}\rangle_{\mathcal{H}} - y_n\right]K_{x_n},$$

where γ is the *step-size*.

Stochastic approximation in the RKHS.

As $\mathcal{R}(g) = \mathbb{E}\left[(\langle g, K_X\rangle_{\mathcal{H}} - Y)^2\right]$, for each pair of observations

$$(\langle g, K_{x_n}\rangle_{\mathcal{H}} - y_n)K_{x_n} = (g(x_n) - y_n)K_{x_n}$$

is an *unbiased stochastic gradient* of $\mathcal{R}$ at g.
Consider the stochastic gradient recursion, starting from $g_0 \in \mathcal{H}$:

$$g_n = g_{n-1} - \gamma\left[\langle g_{n-1}, K_{x_n}\rangle_{\mathcal{H}} - y_n\right]K_{x_n},$$

where γ is the *step-size*. Thus

$$g_n = \sum_{i=1}^{n} a_i K_{x_i},$$

with $(a_n)_{n\geqslant 1}$, $a_n = -\gamma_n(g_{n-1}(x_n) - y_n)$. With averaging,

$$\overline{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} g_k$$

Total complexity: $O(n^2)$

Kernel regression: Analysis

Assume $\mathbb{E}[K(X,X)]$ and $\mathbb{E}\left[Y^2\right]$ are finite. Define the *covariance operator*.

$$\Sigma = \mathbb{E}\left[K_X K_X^\top\right].$$

We make two assumptions:

- Capacity condition: eigenvalue decay of Σ.
- Source condition: position of $g_{\mathcal{H}}$ w.r.t. the kernel space $\mathcal{H}$.

Kernel regression: Analysis

Assume $\mathbb{E}[K(X,X)]$ and $\mathbb{E}\left[Y^2\right]$ are finite. Define the *covariance operator*.

$$\Sigma = \mathbb{E}\left[K_X K_X^\top\right].$$

We make two assumptions:

- Capacity condition: eigenvalue decay of Σ.
- Source condition: position of $g_{\mathcal{H}}$ w.r.t. the kernel space $\mathcal{H}$.

Σ is a trace-class operator, that can be decomposed over its eigen-spaces. Its power: Σ^τ, $\tau > 0.$ are thus well defined.

Capacity condition (CC)

CC(α): for some $\alpha > 1$, we assume that $\mathrm{tr}(\Sigma^{1/\alpha}) < \infty$.

Capacity condition (CC)

CC(α): for some $\alpha > 1$, we assume that $\mathrm{tr}(\Sigma^{1/\alpha}) < \infty$.

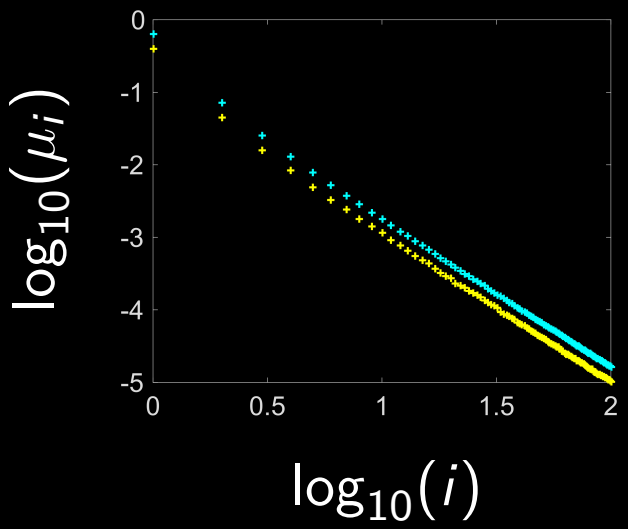
If we denote $(\mu_i)_{i \in I}$ the sequence of non-zero eigenvalues of the operator Σ, in decreasing order, then $\mu_i = O\left(i^{-\alpha}\right)$.

Capacity condition (CC)

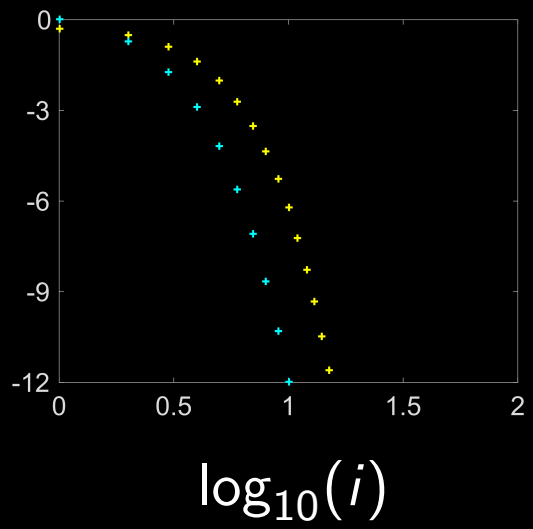
CC(α): for some $\alpha > 1$, we assume that $\mathrm{tr}(\Sigma^{1/\alpha}) < \infty$.

If we denote $(\mu_i)_{i \in I}$ the sequence of non-zero eigenvalues of the operator Σ, in decreasing order, then $\mu_i = O\left(i^{-\alpha}\right)$.

Sobolev first order kernel



Gaussian kernel



Eigenvalue decay of the covariance operator.

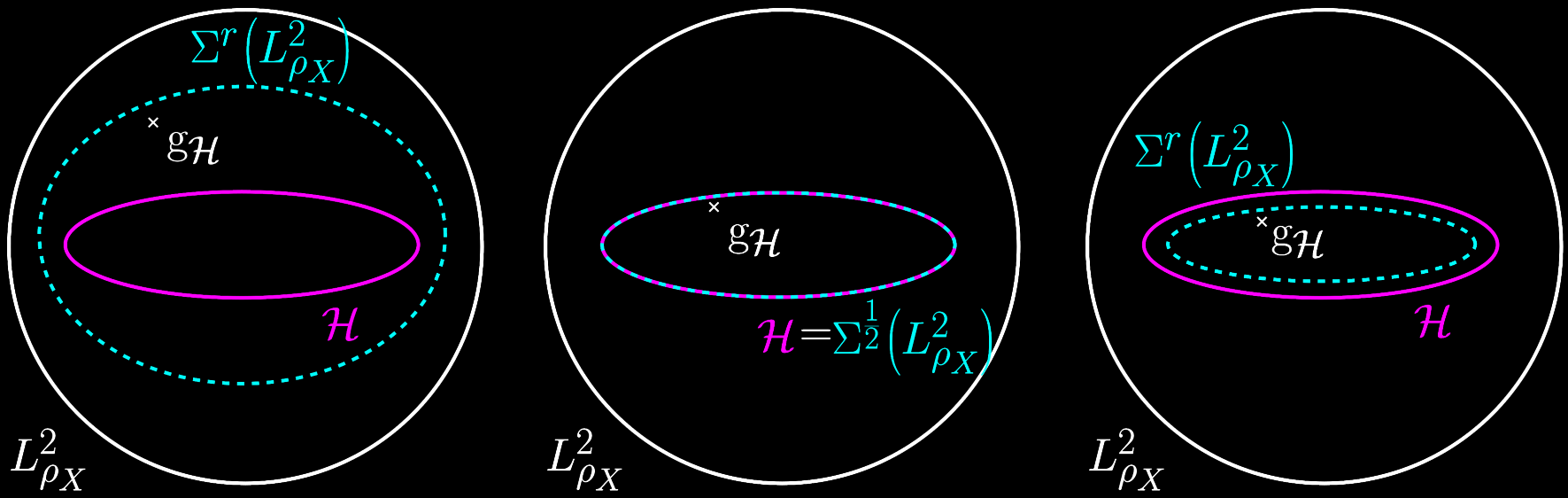
Left: min kernel, $\rho_X = \mathcal{U}[0;1]$, $\longrightarrow$ $CC(\alpha = 2)$.
Right: Gaussian kernel, $\rho_X = \mathcal{U}[-1;1]$. $\longrightarrow$ $CC(\alpha), \forall\ \alpha \geq 1$.

Source condition (SC)

Concerning the optimal function $g_{\mathcal{H}}$, we assume:

SC(r): for some $r \geqslant 0$, $g_{\mathcal{H}} \in \Sigma^r\left(L^2_{\rho_X}\right)$

Thus $\|\Sigma^{-r}(g_{\mathcal{H}})\|_{L^2_{\rho_X}} < \infty$.



NPSA with large step sizes

Theorem

Assume CC(α) and SC(r). Then for any $\gamma \leq \frac{1}{4R^2}$,

$$\mathbb{E}\mathcal{R}(\bar{g}_n) - \mathcal{R}(g_{\mathcal{H}}) \leq \frac{4\sigma^2 \gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}} + 4\frac{\|\Sigma^{-r}(g_{\mathcal{H}} - g_0)\|^2_{L^2_{\rho_X}}}{\gamma^{2r} n^{\min(2r,2)}}.$$

for $\gamma = \gamma_0 n^{\frac{-2\alpha r - 1 + \alpha}{2\alpha r + 1}}$, for $\frac{\alpha - 1}{2\alpha} \leq r \leq 1$

$$\mathbb{E}\mathcal{R}(\bar{g}_n) - \mathcal{R}(g_{\mathcal{H}}) \leq n^{\frac{-2\alpha r}{2\alpha r + 1}} \left(4\sigma^2 \operatorname{tr}(\Sigma^{1/\alpha}) + 4 \left\| \Sigma^{-r}(g_{\mathcal{H}} - g_0) \right\|^2_{L^2_{\rho_X}} \right).$$

NPSA with large step sizes

Theorem

Assume CC(α) and SC(r). Then for any $\gamma \leq \frac{1}{4R^2}$,

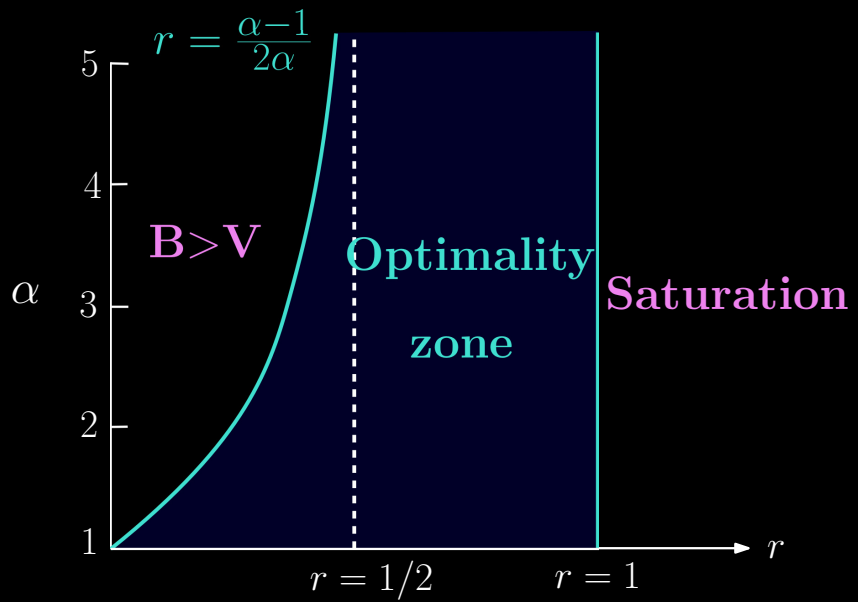
$$\mathbb{E}\mathcal{R}(\bar{g}_n) - \mathcal{R}(g_{\mathcal{H}}) \leq \frac{4\sigma^2 \gamma^{1/\alpha} \operatorname{tr}(\Sigma^{1/\alpha})}{n^{1-1/\alpha}} + 4\frac{\|\Sigma^{-r}(g_{\mathcal{H}} - g_0)\|^2_{L^2_{\rho_X}}}{\gamma^{2r} n^{\min(2r,2)}}.$$

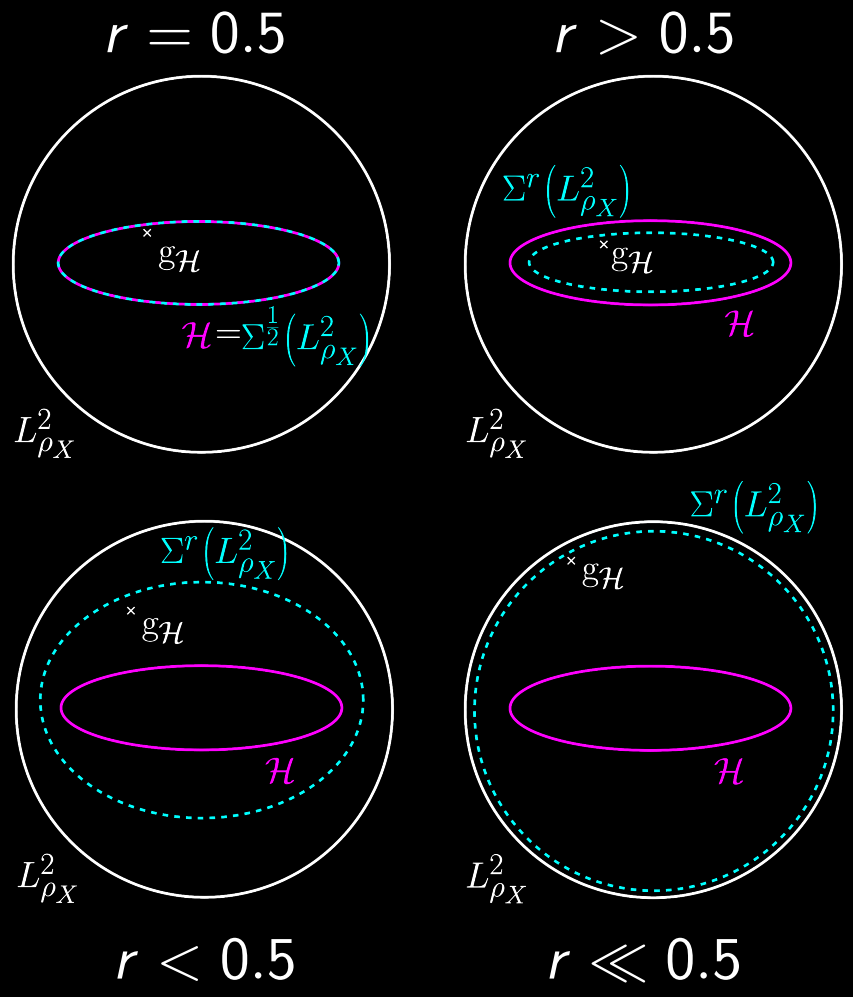
for $\gamma = \gamma_0 n^{\frac{-2\alpha r - 1 + \alpha}{2\alpha r + 1}}$, for $\frac{\alpha - 1}{2\alpha} \leq r \leq 1$

$$\mathbb{E}\mathcal{R}(\bar{g}_n) - \mathcal{R}(g_{\mathcal{H}}) \leq n^{\frac{-2\alpha r}{2\alpha r + 1}} \left(4\sigma^2 \operatorname{tr}(\Sigma^{1/\alpha}) + 4 \left\|\Sigma^{-r}(g_{\mathcal{H}} - g_0)\right\|^2_{L^2_{\rho_X}} \right).$$

- Statistically optimal rate. [Caponnetto and De Vito, 2007].
- *Beyond:* online, minimal assumptions...

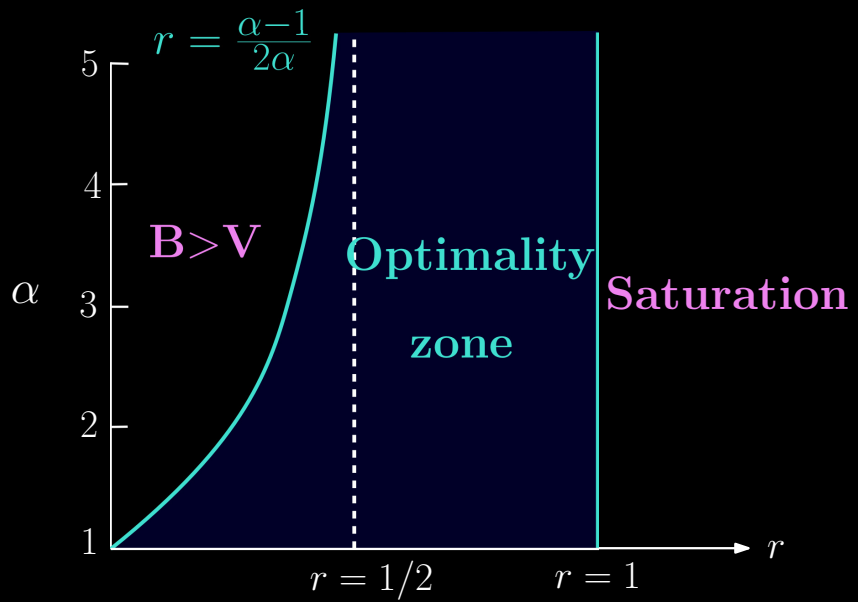
Optimality regions

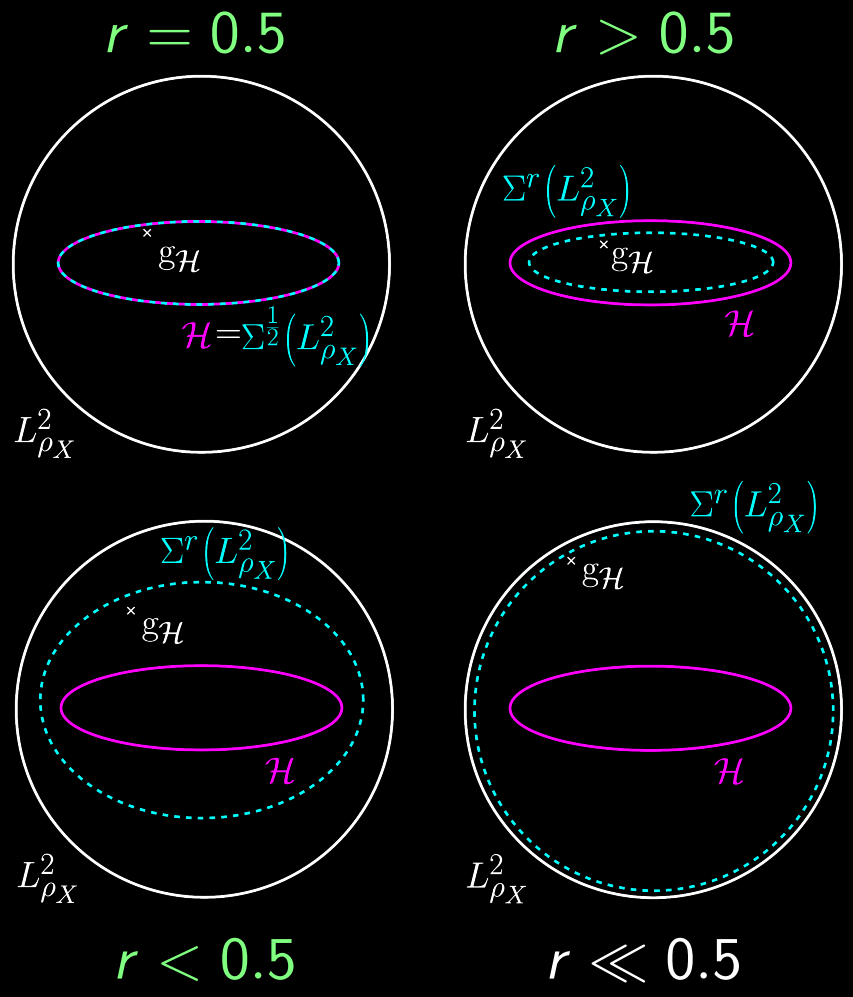




Optimal rate in RKHS can be achieved via large step size and averaging in many situations.

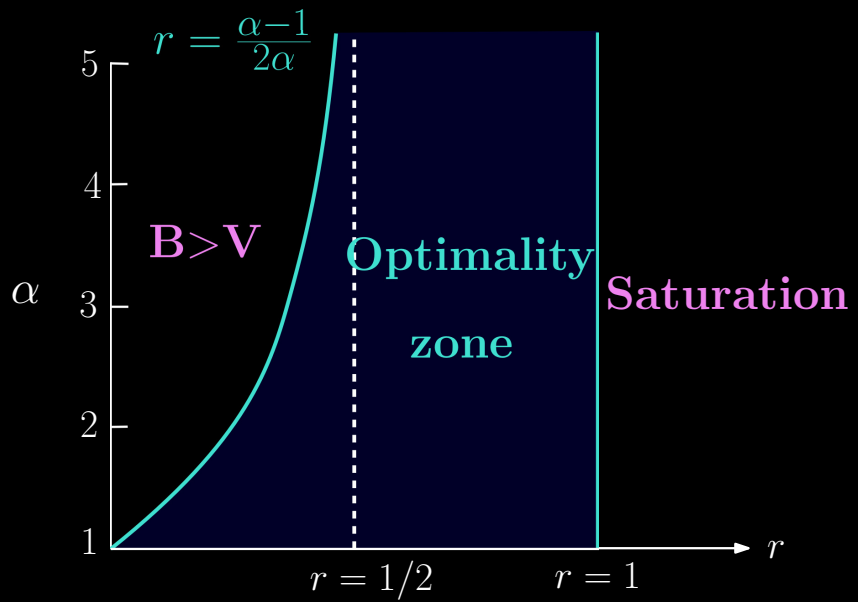
Optimality regions

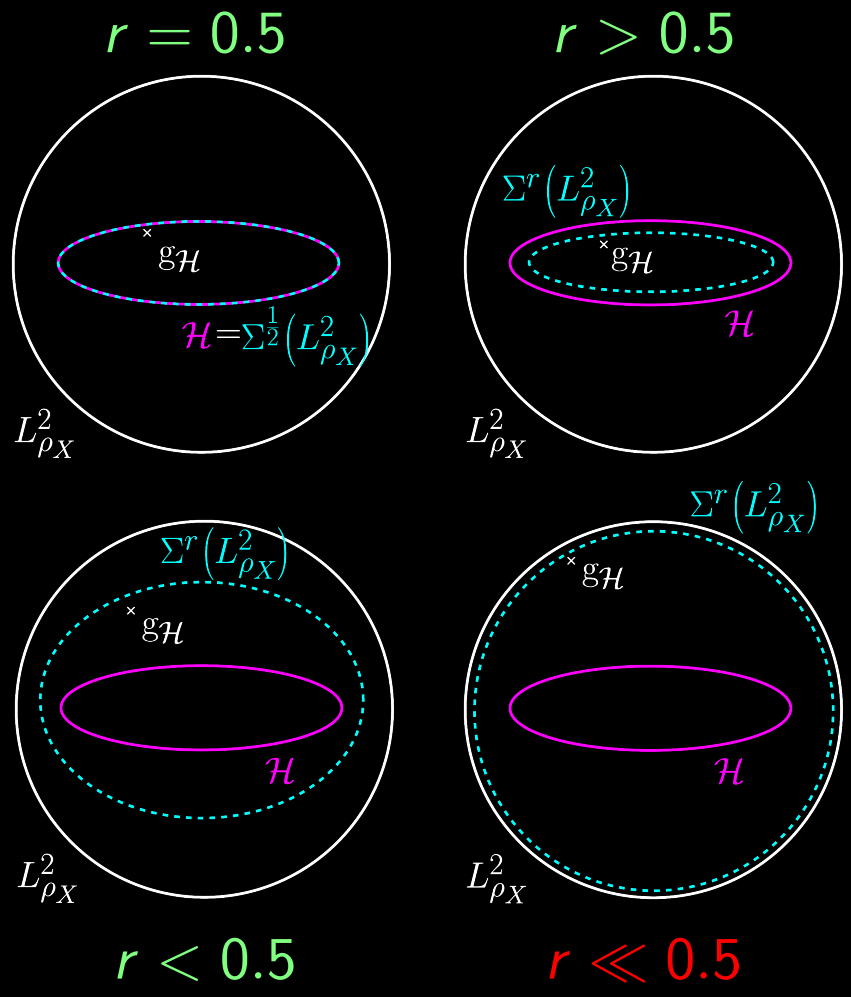




Optimal rate in RKHS can be achieved via large step size and averaging in many situations.

Optimality regions





Optimal rate in RKHS can be achieved via large step size and averaging in many situations.

Acceleration: Reproducing kernel Hilbert space setting

We consider the RKHS setting presented before.

Theorem

Assume CC(α) and SC(r). Then for $\gamma = \gamma_0 n^{-\frac{4r\alpha+2-\alpha}{2r\alpha+1}}$, for $\lambda = \frac{1}{\gamma n^2}$, for $r \geq \frac{\alpha-2}{2\alpha}$,

$$\mathbb{E}\mathcal{R}\left(\bar{g}_n\right) - \mathcal{R}(g_{\mathcal{H}}) \leq C_{\theta_0,\theta_*,\Sigma} \quad n^{\frac{-2\alpha r}{2\alpha r+1}} \ .$$

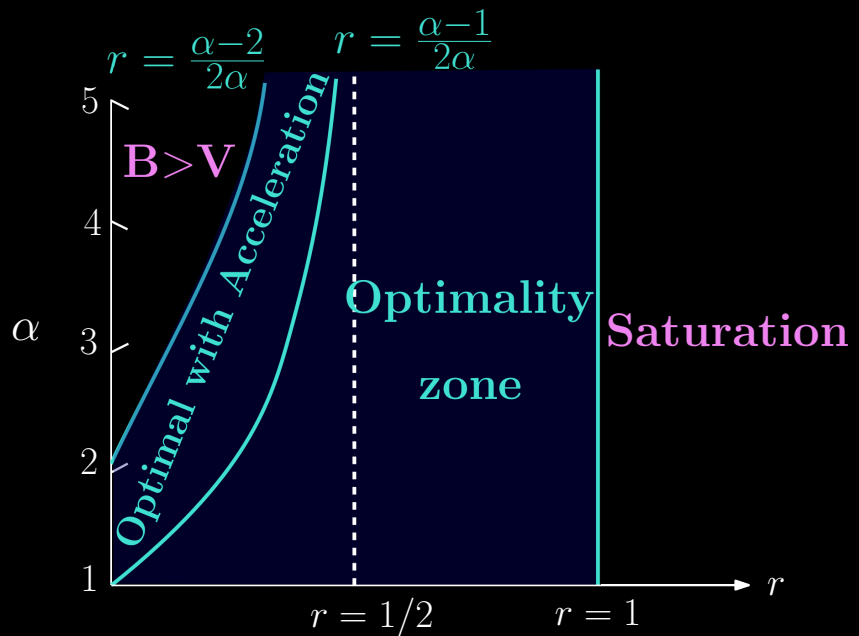
Acceleration: Reproducing kernel Hilbert space setting

We consider the RKHS setting presented before.

Theorem

Assume CC(α) and SC(r). Then for $\gamma = \gamma_0 n^{-\frac{4r\alpha+2-\alpha}{2r\alpha+1}}$, for $\lambda = \frac{1}{\gamma n^2}$, for $r \geq \frac{\alpha-2}{2\alpha}$,

$$\mathbb{E}\mathcal{R}\left(\bar{g}_n\right) - \mathcal{R}(g_{\mathcal{H}}) \leq C_{\theta_0,\theta_*,\Sigma} \quad n^{\frac{-2\alpha r}{2\alpha r+1}} .$$



Least squares: some conclusions

- Provide optimal rate of convergence under two assumptions for non-parametric regression in Hilbert spaces: large step sizes and averaging.

Least squares: some conclusions

- Provide optimal rate of convergence under two assumptions for non-parametric regression in Hilbert spaces: large step sizes and averaging.
- Sheds some light on FD case.

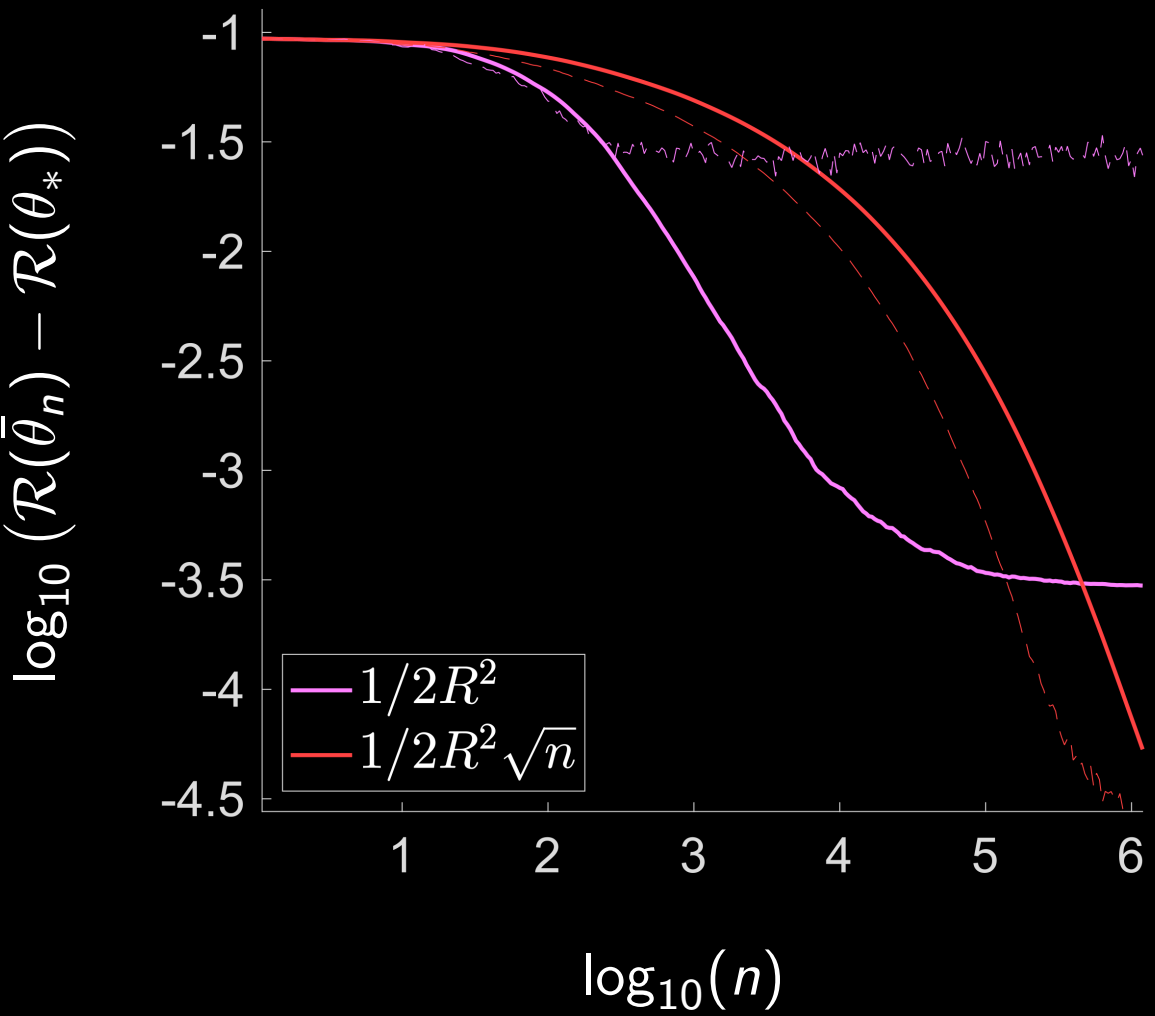
Least squares: some conclusions

- Provide optimal rate of convergence under two assumptions for non-parametric regression in Hilbert spaces: large step sizes and averaging.
- Sheds some light on FD case.
- Possible to attain simultaneously optimal rate from the statistical and optimization point of view.

Outline

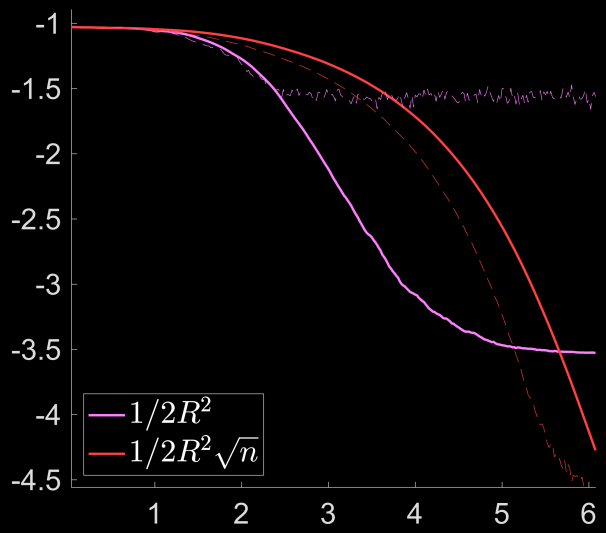
1. Introduction.
2. Non-parametric stochastic approximation
3. Faster rates with acceleration
4. Stochastic approximation as a Markov chain: extension to non quadratic loss functions.
   - Motivation
   - Assumptions
   - Convergence in Wasserstein distance.

Motivation 1/ 2. Large step sizes!



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

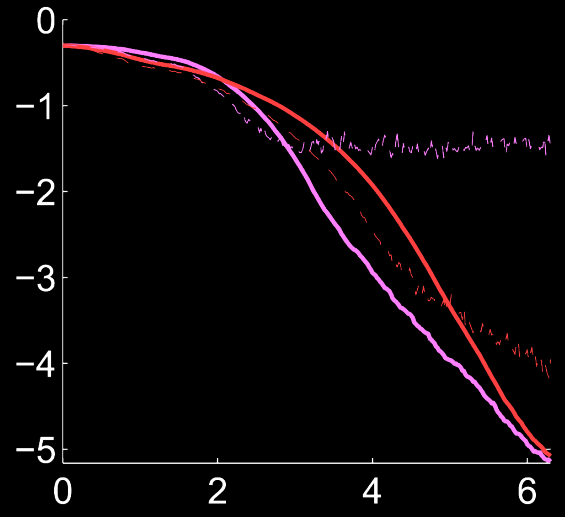
Motivation 2/ 2. Difference between quadratic and logistic loss



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: 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:

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

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

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.**

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:

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

[†] *Dieuleveut, Durmus, Bach [2017].*

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,

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

[†] *Dieuleveut, Durmus, Bach [2017].*

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

[†] *Dieuleveut, Durmus, Bach [2017].*

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

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

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

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

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

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

Wasserstein metric: distance between probability measures.

Behavior under limit distribution.

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

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 $\theta_1 \sim \pi_\gamma$.

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

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

Behavior under limit distribution.

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

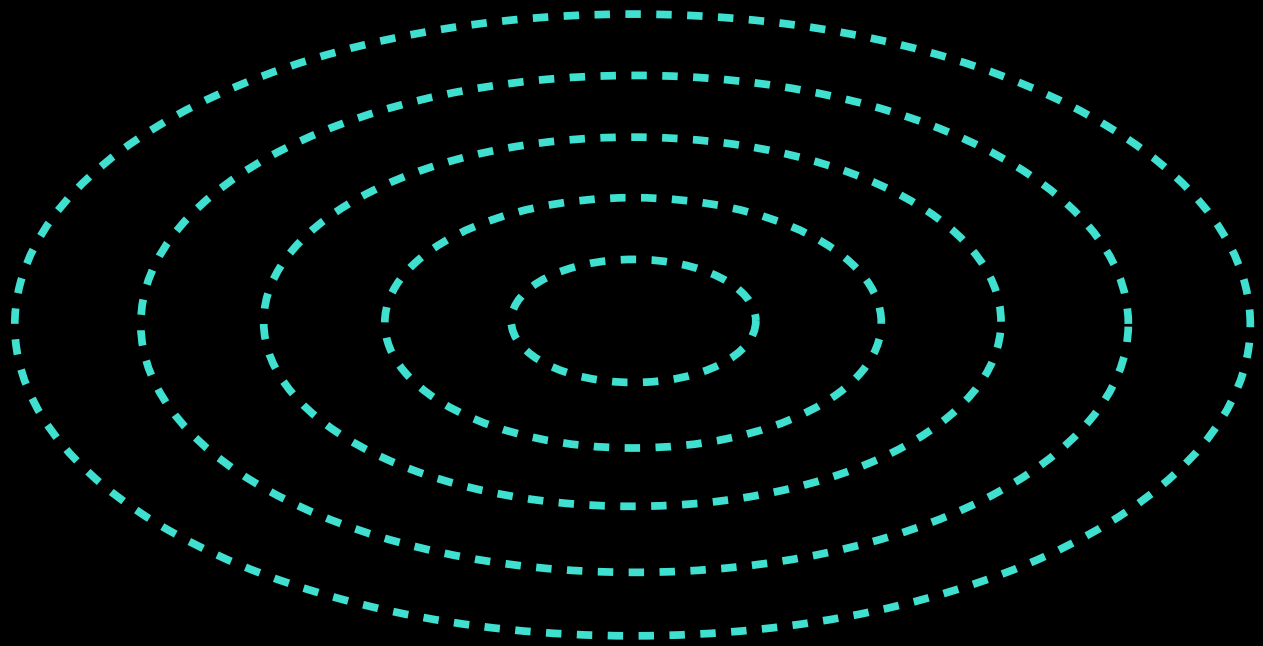
If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.

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

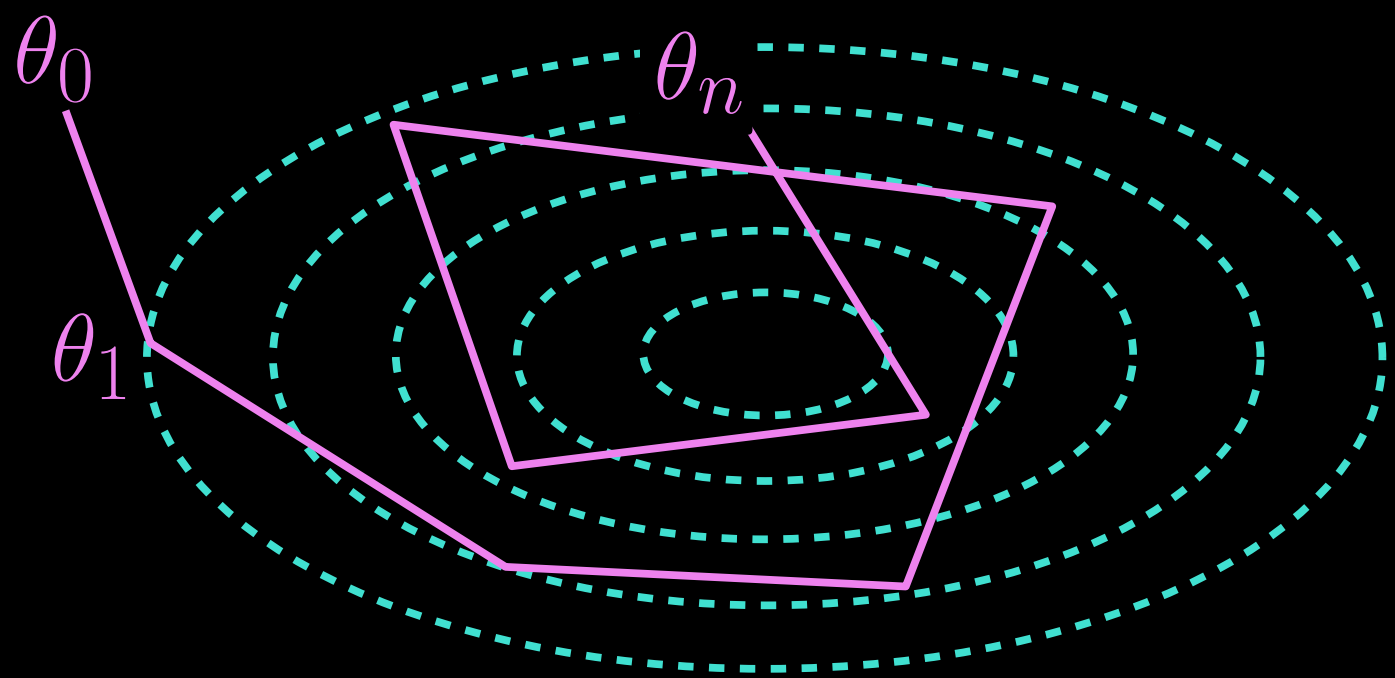
$$\mathbb{E}_{\pi_\gamma}\left[\mathcal{R}'(\theta)\right] = 0$$

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

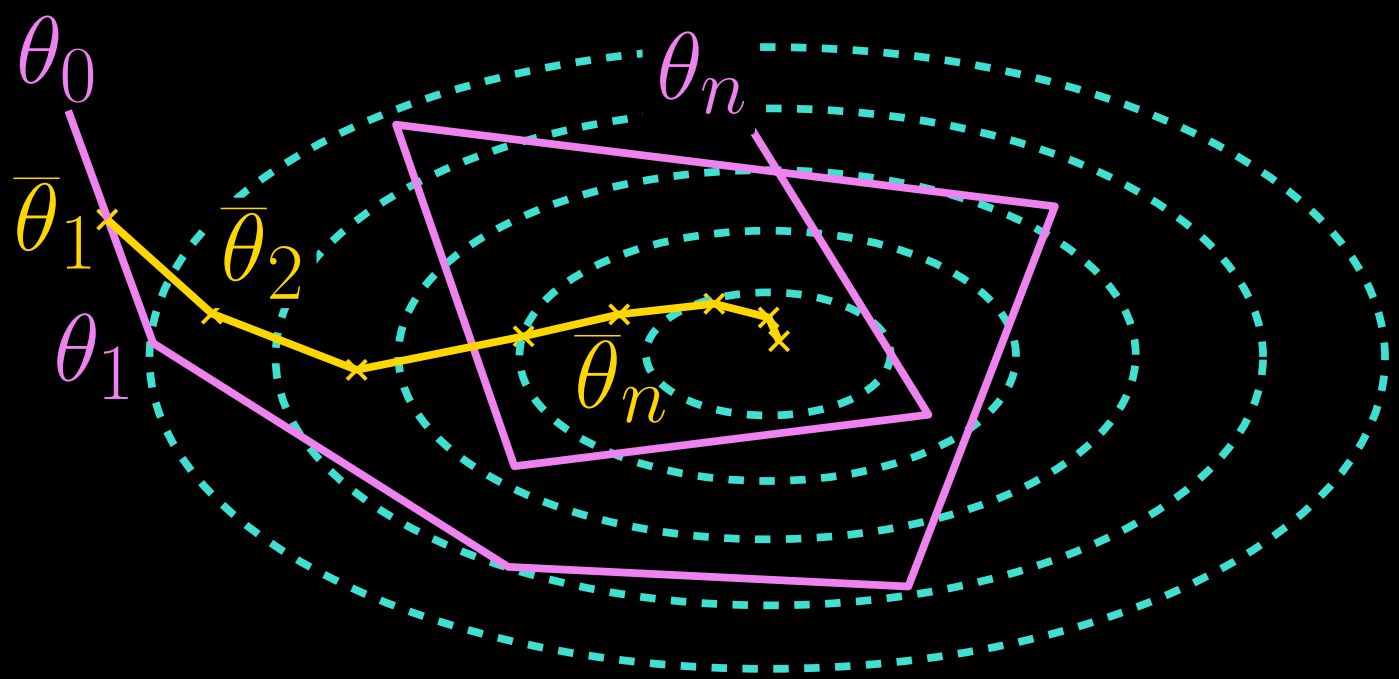
Constant learning rate SGD: convergence in the quadratic case



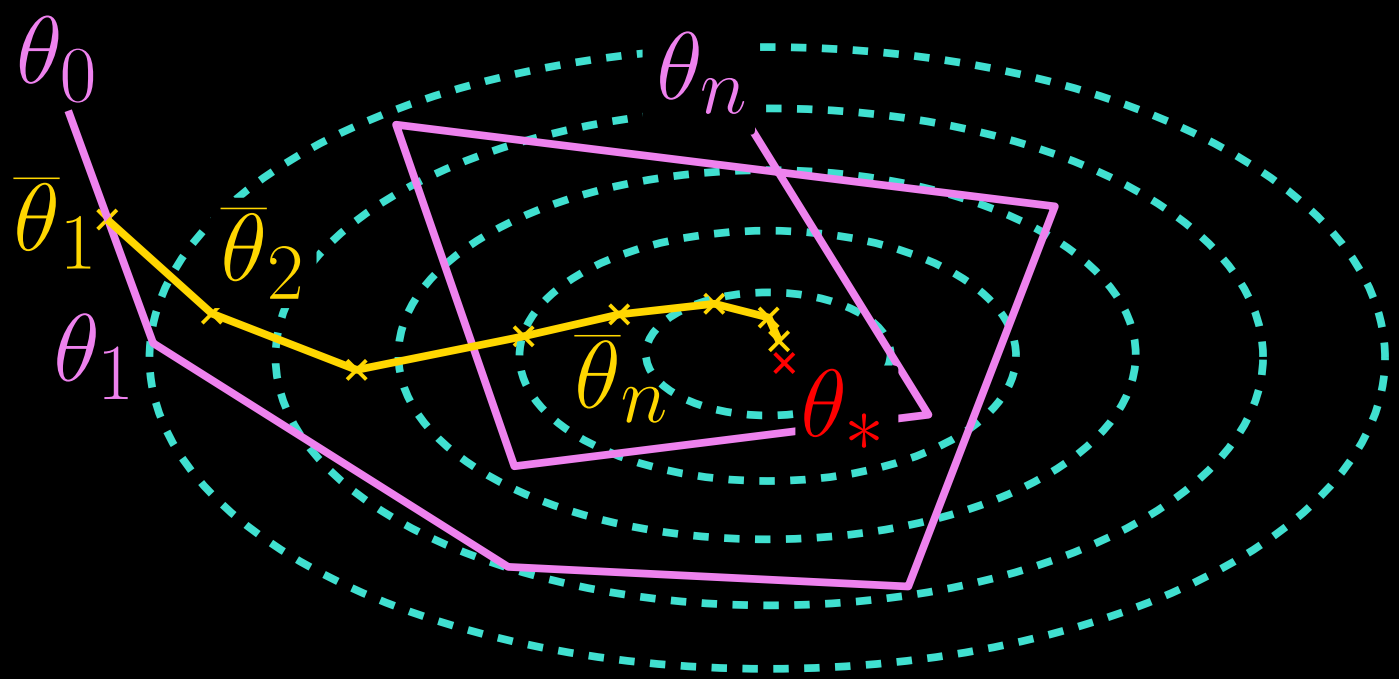
Constant learning rate SGD: convergence in the quadratic case



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 $\theta_1 \sim \pi_\gamma$.

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

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

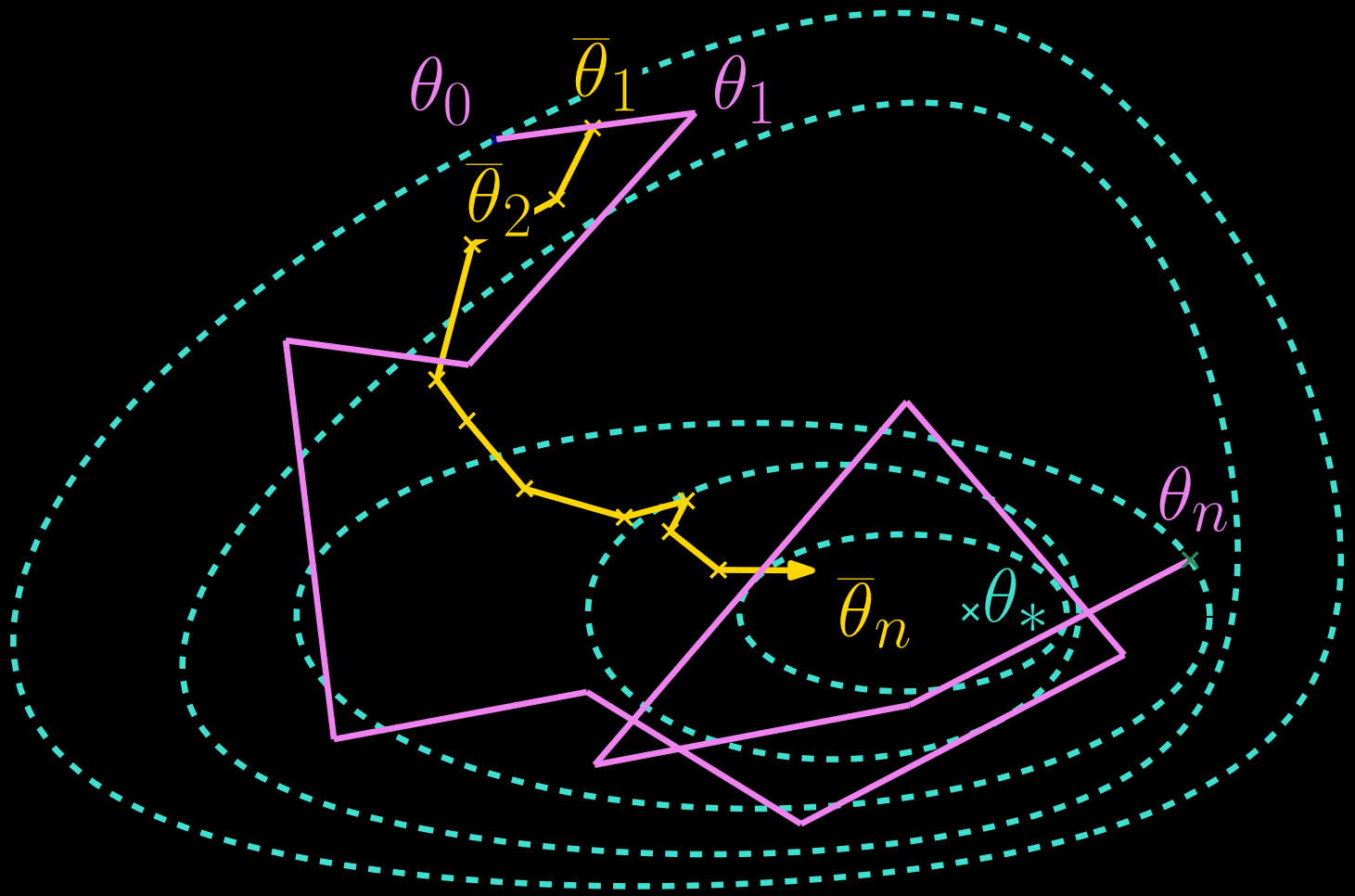
In the quadratic case (linear gradients) $\Sigma\mathbb{E}_{\pi_\gamma}[\theta - \theta_*] = 0$: $\bar{\theta}_\gamma = \theta_*$!

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

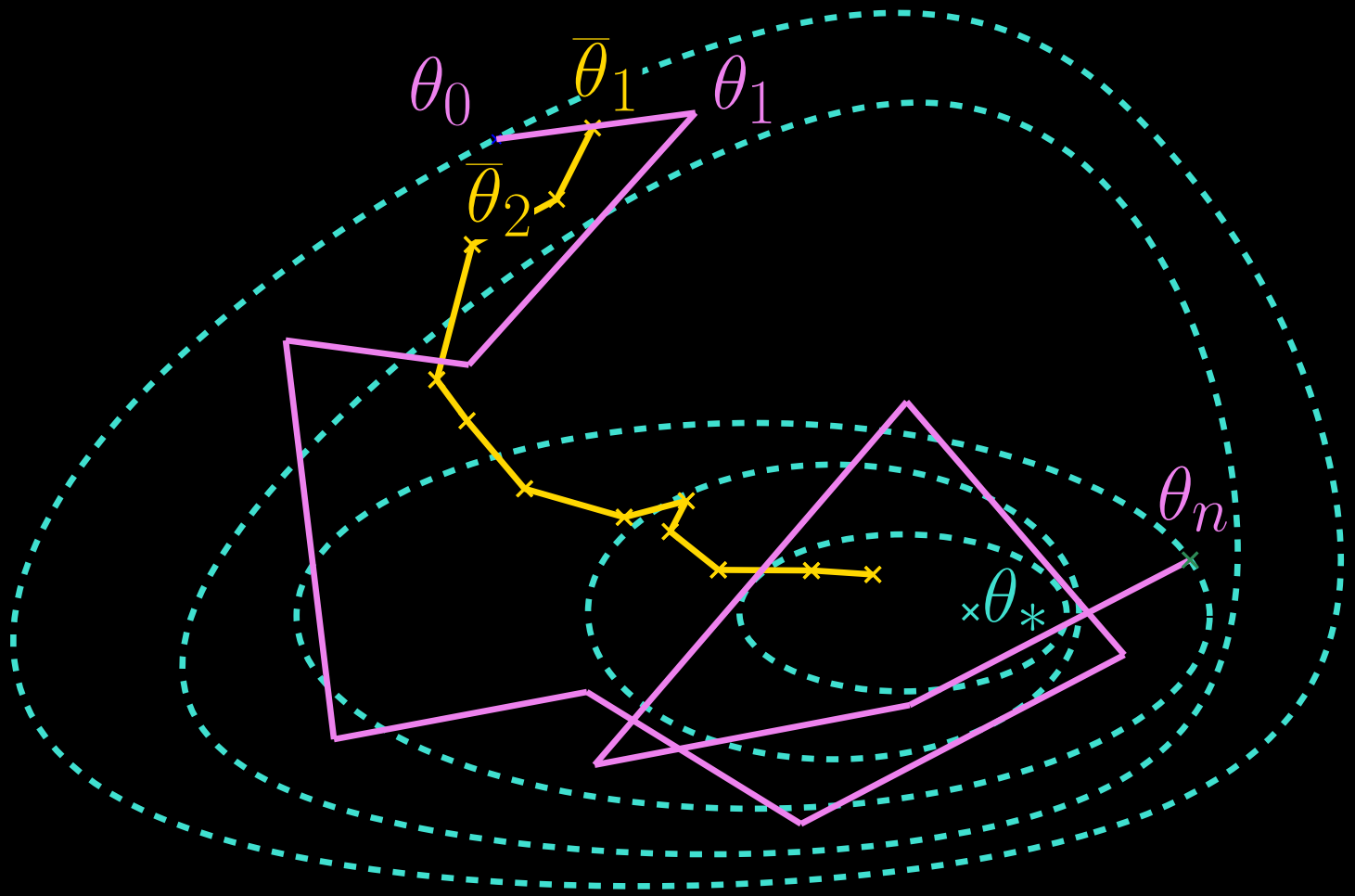
$$\bar{\theta}_\gamma - \theta_* = \gamma\mathcal{R}''(\theta_*)^{-1}\mathcal{R}'''(\theta_*)\Big(\left[\mathcal{R}''(\theta_*) \otimes I + I \otimes \mathcal{R}''(\theta_*)\right]^{-1}\mathbb{E}_{\pi_\gamma,\varepsilon}[\varepsilon(\theta)^{\otimes 2}]\Big) + O(\gamma^2)$$

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

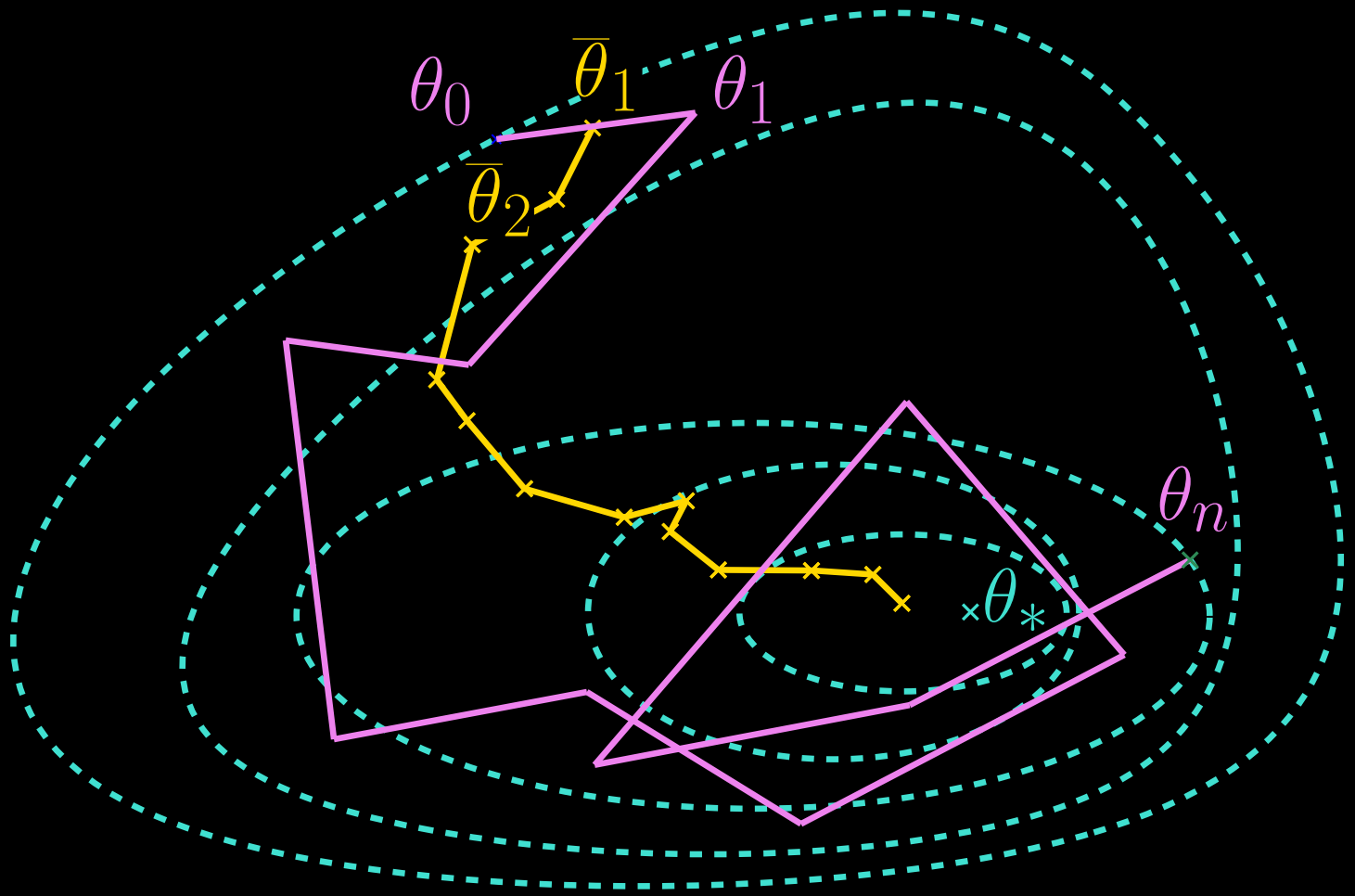
Constant learning rate SGD: convergence in the non-quadratic case



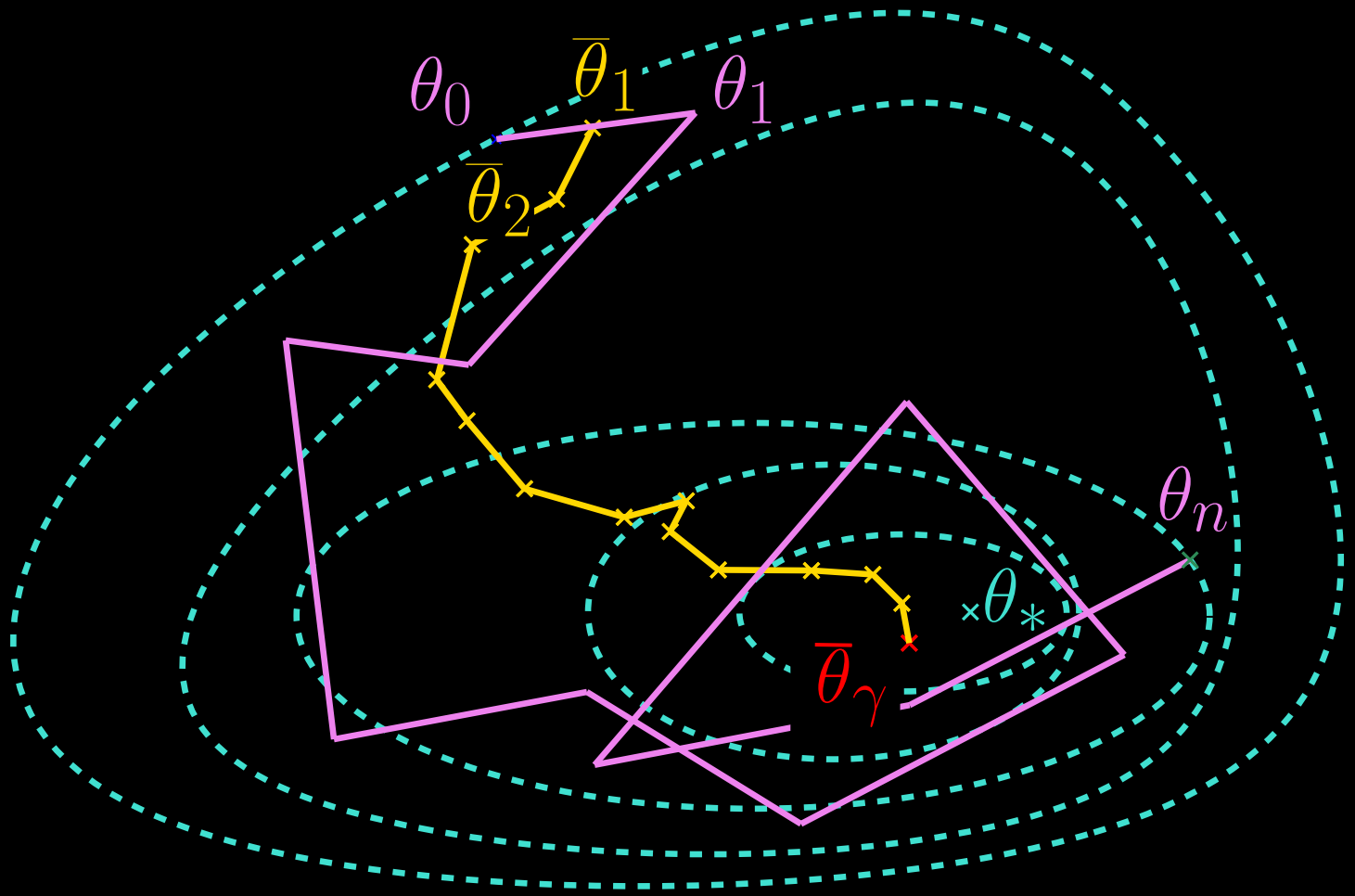
Constant learning rate SGD: convergence in the non-quadratic case



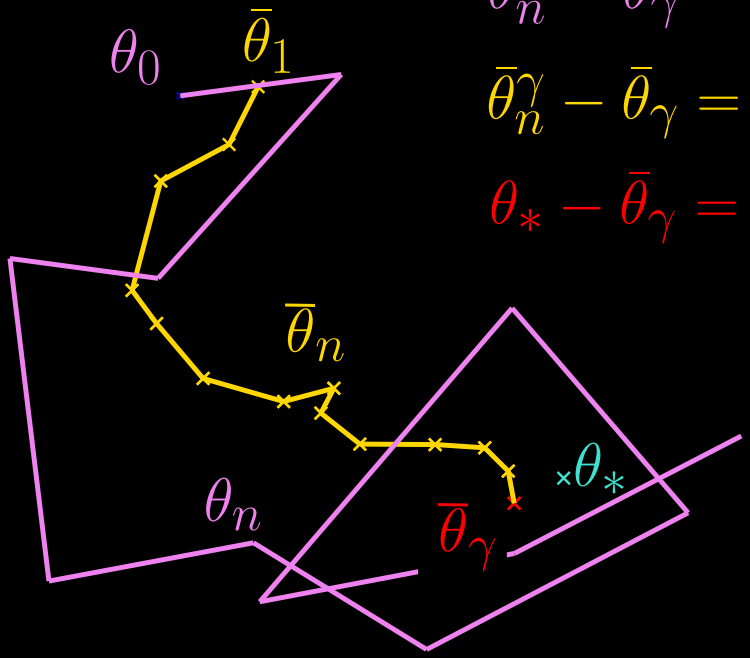
Constant learning rate SGD: convergence in the non-quadratic case



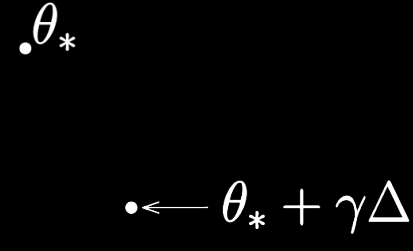
Constant learning rate SGD: convergence in the non-quadratic case



Richardson extrapolation



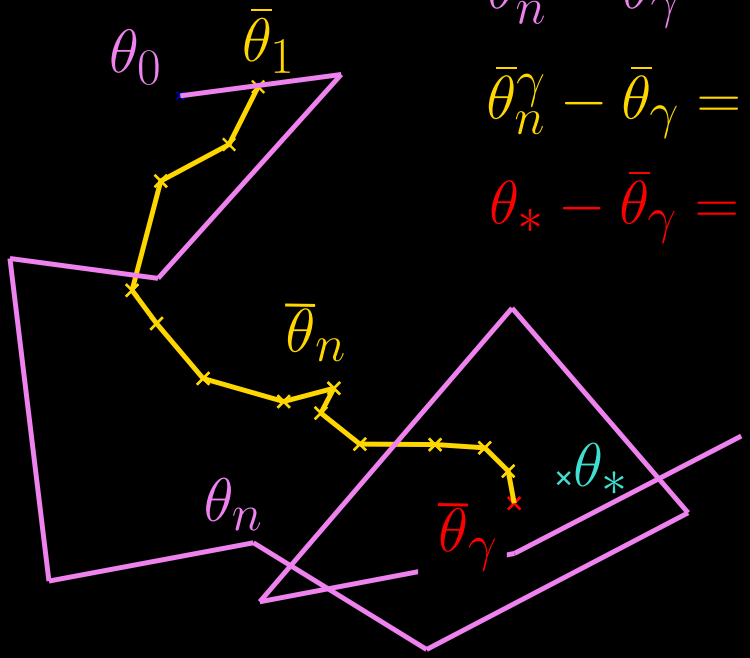
$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



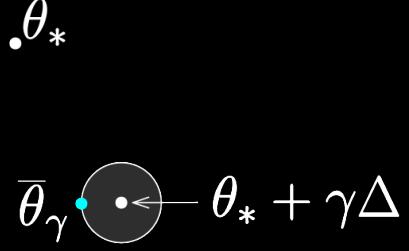
Recovering convergence closer to θ_* by **Richardson extrapolation**

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

Richardson extrapolation



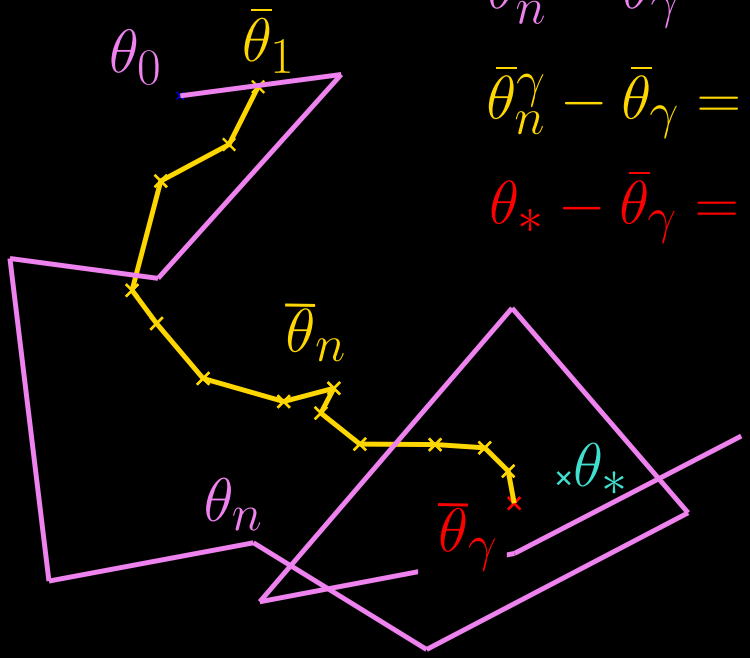
$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



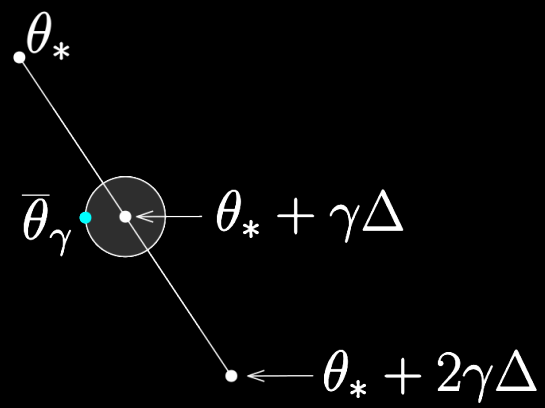
Recovering convergence closer to θ_* by **Richardson extrapolation**

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

Richardson extrapolation



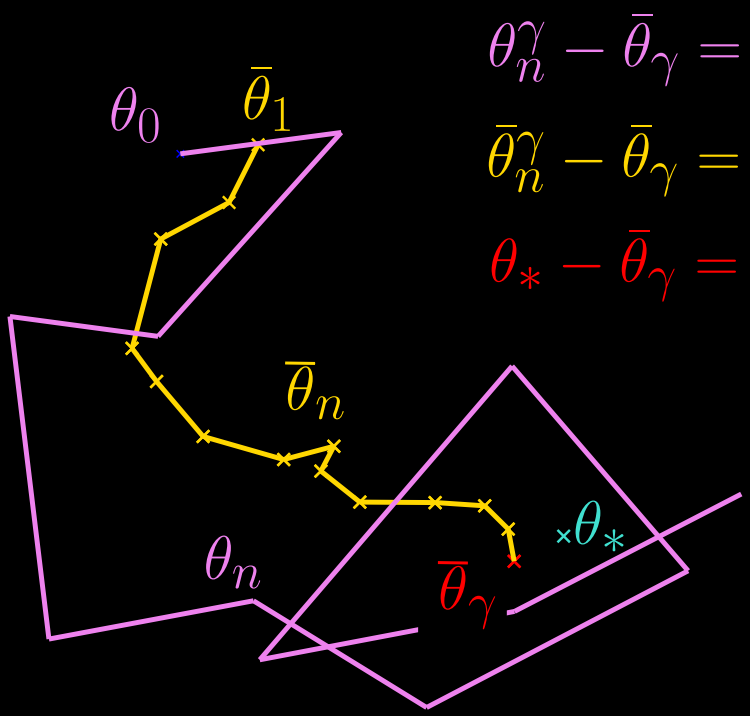
$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



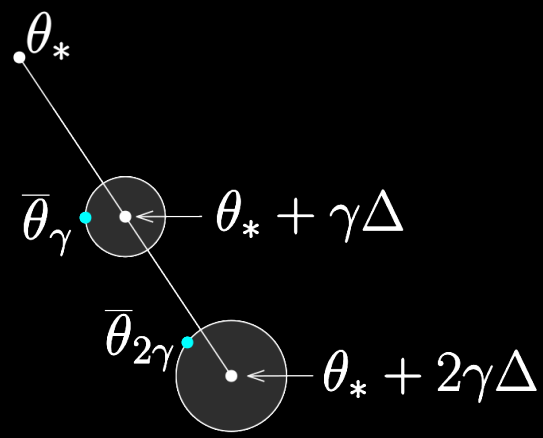
Recovering convergence closer to θ_* by **Richardson extrapolation**

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

Richardson extrapolation



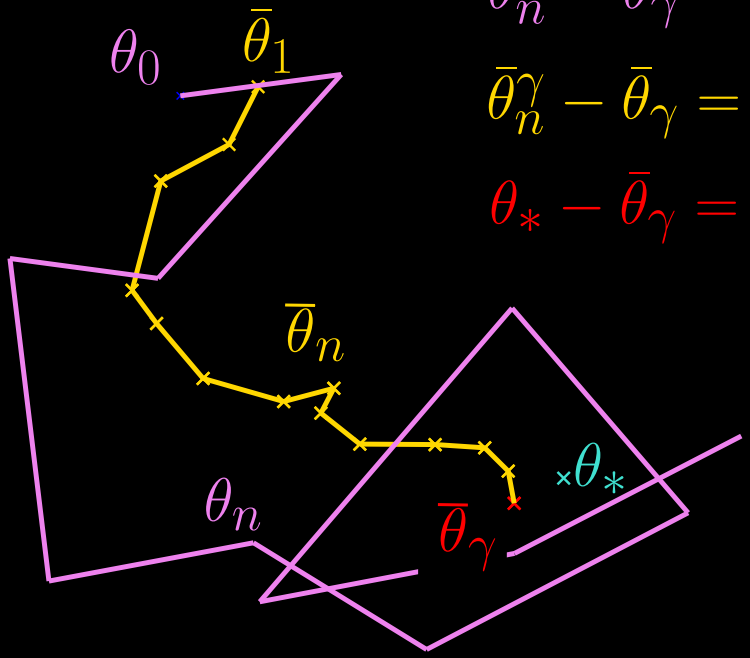
$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



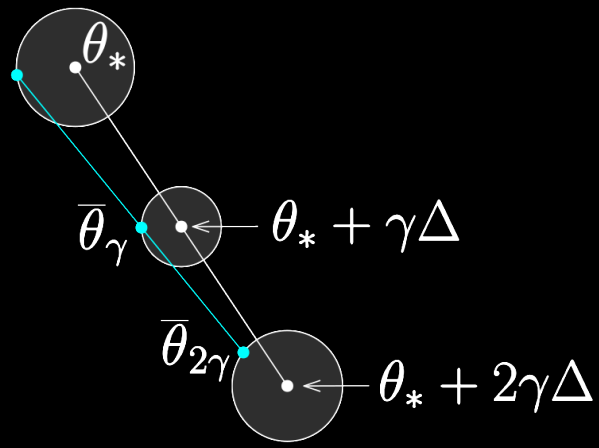
Recovering convergence closer to θ_* by **Richardson extrapolation**

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

Richardson extrapolation



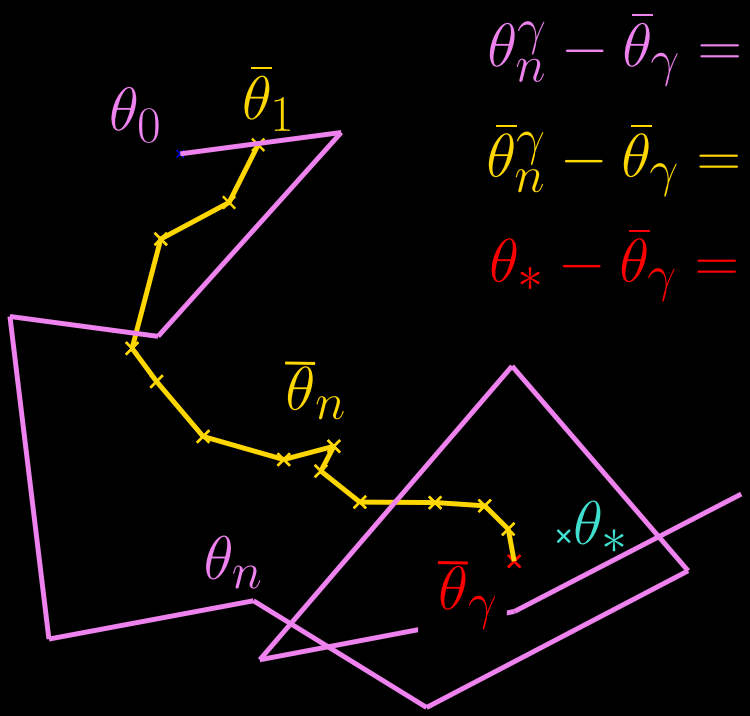
$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



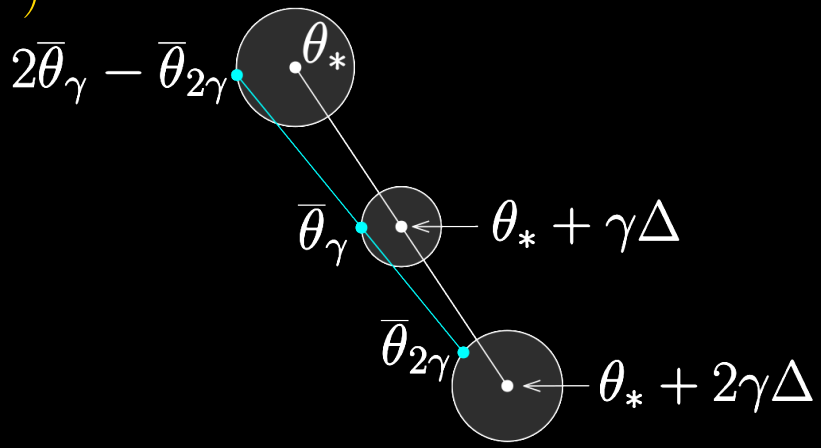
Recovering convergence closer to θ_* by **Richardson extrapolation**

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

Richardson extrapolation



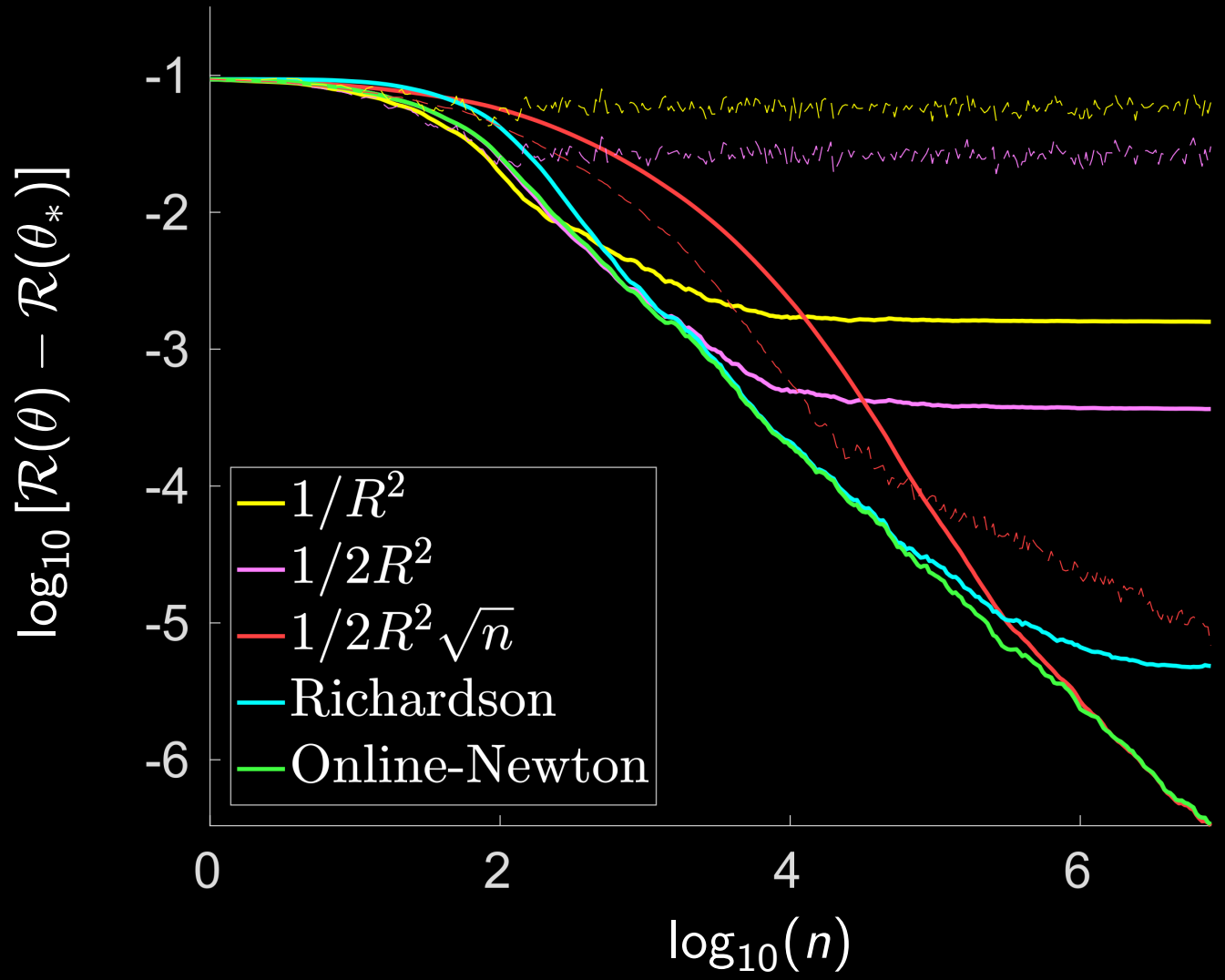
$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



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

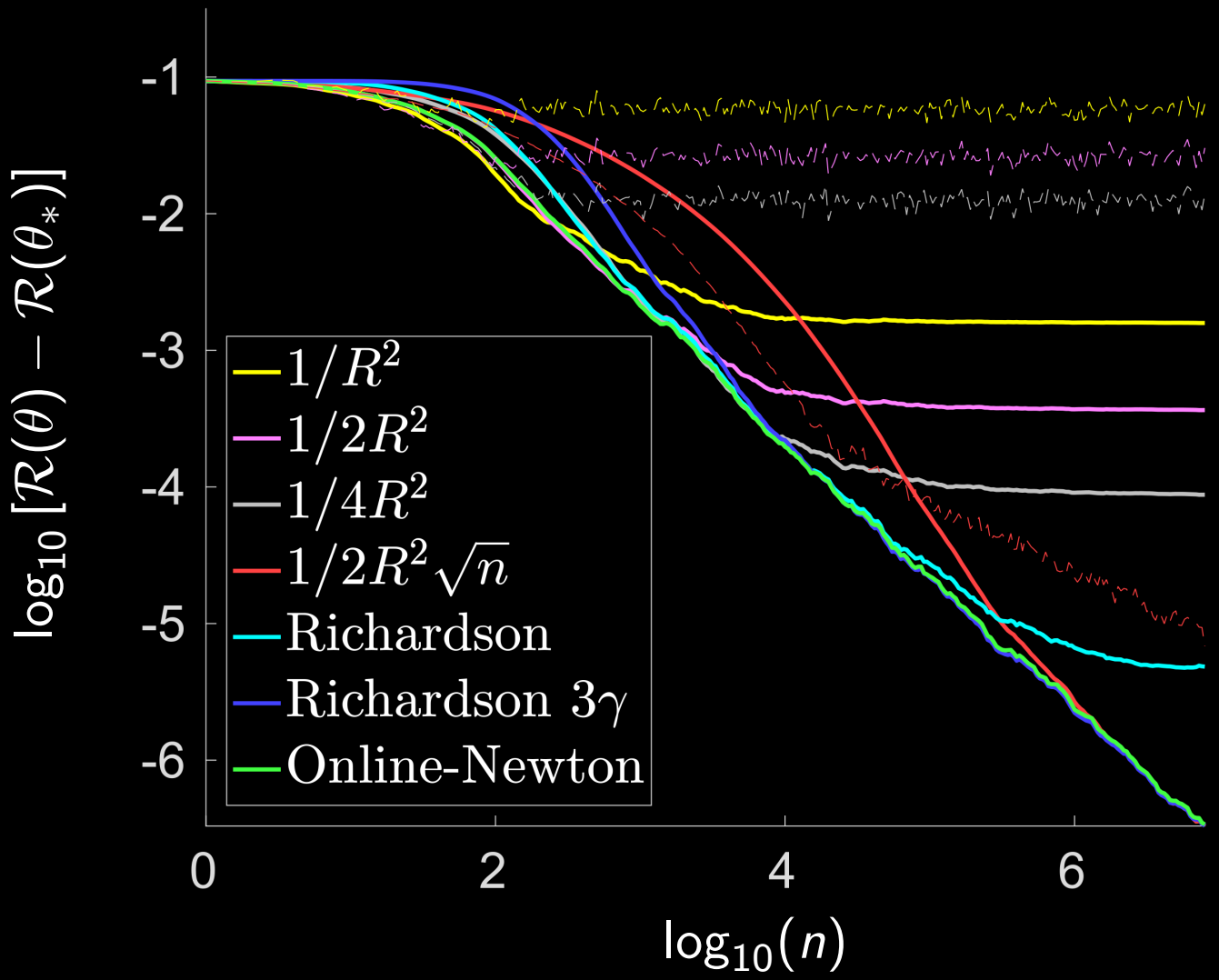
$$2\bar{\theta}_{n,\gamma} - \bar{\theta}_{n,2\gamma}$$

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 message:

- Precise description of the 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.
- Extrapolation tricks can help.
- *Beyond:* new error decomposition (link with diffusions), ...

Open directions

- *Markov chain, beyond strong convexity*

Open directions

- *Markov chain, beyond strong convexity*
- *Adaptivity for non-parametric regression*

Open directions

- *Markov chain, beyond strong convexity*
- *Adaptivity for non-parametric regression*
- *Complexity of non-parametric regression. Stochastic gradient descent and random features*.

Open directions

- *Markov chain, beyond strong convexity*
- *Adaptivity for non-parametric regression*
- *Complexity of non-parametric regression. Stochastic gradient descent and random features*.
- *Density estimation*.

[noframenumbering]

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

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

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183, 2008.

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

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

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

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

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

P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths. *EEE Transactions in Information Theory*, (99):5716–5735, 2011.

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 2008.