

# Statistical machine learning and convex optimization

**Francis Bach**

*INRIA - Ecole Normale Supérieure, Paris, France*



StatMathAppli 2017, Fréjus - September 2017

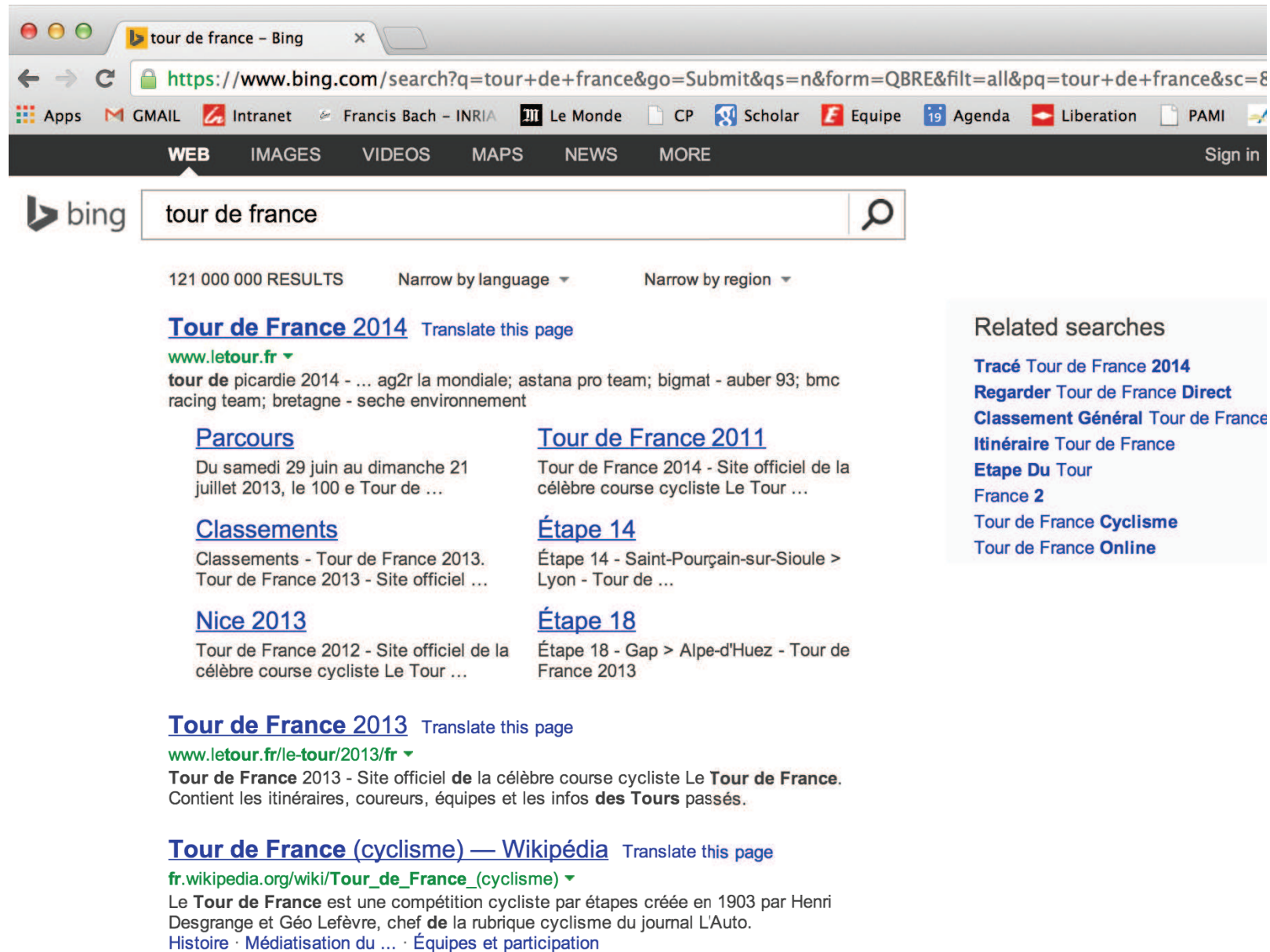
Slides available: [www.di.ens.fr/~fbach/fbach\\_frejus\\_2017.pdf](http://www.di.ens.fr/~fbach/fbach_frejus_2017.pdf)

# “Big data” revolution?

## A new scientific context

- **Data everywhere:** size does not (always) matter
- **Science and industry**
- **Size and variety**
- **Learning from examples**
  - $n$  observations in dimension  $d$

# Search engines - Advertising



The image shows a screenshot of a web browser displaying a Bing search results page for the query "tour de france". The browser's address bar shows the URL: <https://www.bing.com/search?q=tour+de+france&go=Submit&qsn=n&form=QBRE&filt=all&pq=tour+de+france&sc=8>. The search bar contains the text "tour de france".

The search results show 121 000 000 results. The top result is "Tour de France 2014" from [www.letour.fr](http://www.letour.fr). The snippet for this result reads: "tour de picardie 2014 - ... ag2r la mondiale; astana pro team; bigmat - auber 93; bmc racing team; bretagne - seche environnement".

Below the top result, there are several sub-sections:

- Parcours**: Du samedi 29 juin au dimanche 21 juillet 2013, le 100 e Tour de ...
- Classements**: Classements - Tour de France 2013. Tour de France 2013 - Site officiel ...
- Nice 2013**: Tour de France 2012 - Site officiel de la célèbre course cycliste Le Tour ...
- Tour de France 2011**: Tour de France 2014 - Site officiel de la célèbre course cycliste Le Tour ...
- Étape 14**: Étape 14 - Saint-Pourçain-sur-Sioule > Lyon - Tour de ...
- Étape 18**: Étape 18 - Gap > Alpe-d'Huez - Tour de France 2013

On the right side of the page, there is a "Related searches" section with the following links:

- Tracé Tour de France 2014
- Regarder Tour de France Direct
- Classement Général Tour de France
- Itinéraire Tour de France
- Étape Du Tour
- France 2
- Tour de France Cyclisme
- Tour de France Online

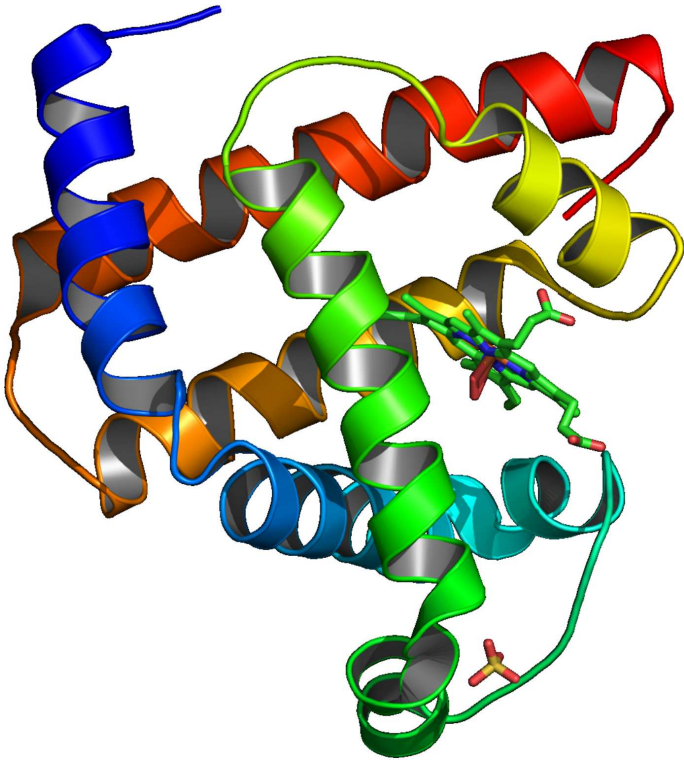
At the bottom of the page, there is another result for "Tour de France 2013" from [www.letour.fr/le-tour/2013/fr](http://www.letour.fr/le-tour/2013/fr). The snippet reads: "Tour de France 2013 - Site officiel de la célèbre course cycliste Le Tour de France. Contient les itinéraires, coureurs, équipes et les infos des Tours passés."

Finally, there is a result for "Tour de France (cyclisme) — Wikipédia" from [fr.wikipedia.org/wiki/Tour\\_de\\_France\\_\(cyclisme\)](http://fr.wikipedia.org/wiki/Tour_de_France_(cyclisme)). The snippet reads: "Le Tour de France est une compétition cycliste par étapes créée en 1903 par Henri Desgrange et Géo Lefèvre, chef de la rubrique cyclisme du journal L'Auto. Histoire · Médiatisation du ... · Équipes et participation".

# Visual object recognition



# Bioinformatics



- **Protein:** Crucial elements of cell life
- **Massive data:** 2 millions for humans
- **Complex data**

# Context

## Machine learning for “big data”

- **Large-scale machine learning:** **large  $d$ , large  $n$** 
  - $d$  : dimension of each observation (input)
  - $n$  : number of observations
- **Examples:** computer vision, bioinformatics, advertising

# Context

## Machine learning for “big data”

- **Large-scale machine learning:** **large  $d$ , large  $n$** 
  - $d$  : dimension of each observation (input)
  - $n$  : number of observations
- **Examples:** computer vision, bioinformatics, advertising
- **Ideal running-time complexity:**  $O(dn)$

# Context

## Machine learning for “big data”

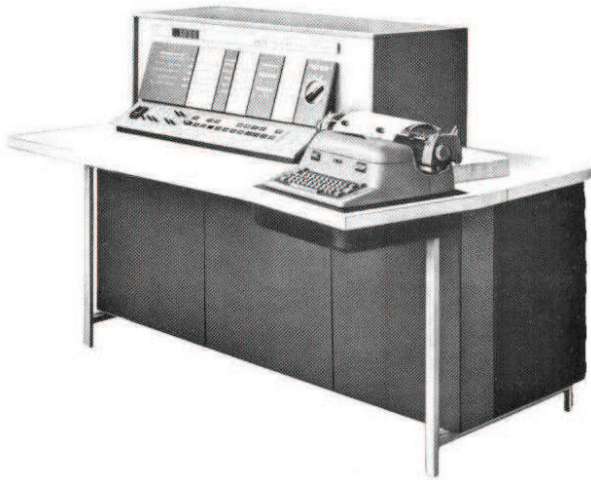
- **Large-scale machine learning:** **large  $d$ , large  $n$** 
  - $d$  : dimension of each observation (input)
  - $n$  : number of observations
- **Examples:** computer vision, bioinformatics, advertising
- **Ideal running-time complexity:**  $O(dn)$
- **Going back to simple methods**
  - Stochastic gradient methods (Robbins and Monro, 1951b)
  - Mixing statistics and optimization



# Scaling to large problems

## “Retour aux sources”

- **1950's:** Computers not powerful enough



IBM “1620”, 1959

CPU frequency: 50 KHz

Price > 100 000 dollars

- **2010's:** Data too massive

# Scaling to large problems

## “Retour aux sources”

- **1950's**: Computers not powerful enough



IBM “1620”, 1959

CPU frequency: 50 KHz

Price > 100 000 dollars

- **2010's**: Data too massive
- **Stochastic gradient methods** (Robbins and Monro, 1951a)
  - Going back to simple methods

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

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

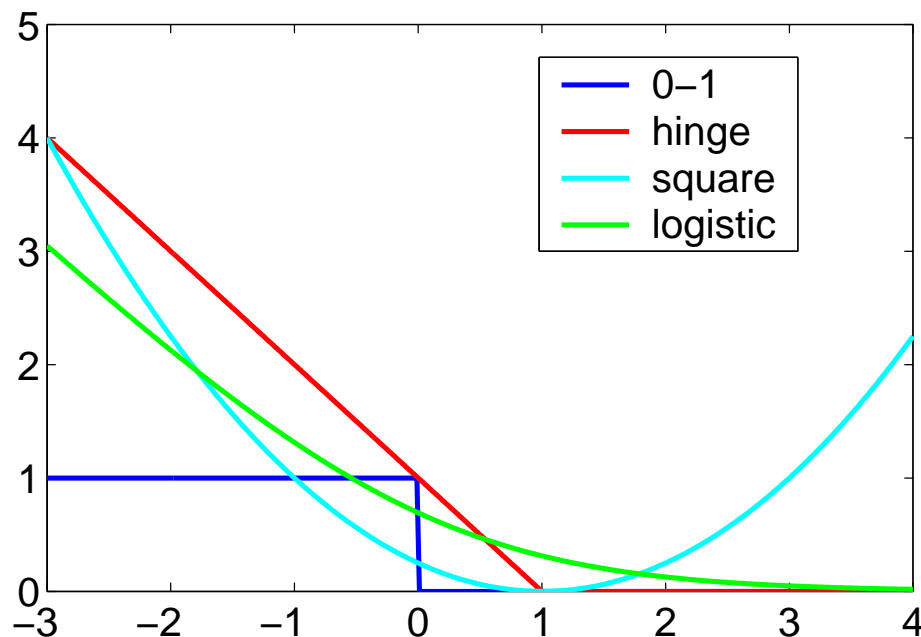
convex data fitting term + regularizer

## Usual losses

- **Regression:**  $y \in \mathbb{R}$ , prediction  $\hat{y} = \theta^\top \Phi(x)$ 
  - quadratic loss  $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \theta^\top \Phi(x))^2$

# Usual losses

- **Regression:**  $y \in \mathbb{R}$ , prediction  $\hat{y} = \theta^\top \Phi(x)$ 
  - quadratic loss  $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \theta^\top \Phi(x))^2$
- **Classification :**  $y \in \{-1, 1\}$ , prediction  $\hat{y} = \text{sign}(\theta^\top \Phi(x))$ 
  - loss of the form  $\ell(y \theta^\top \Phi(x))$
  - “True” **0-1** loss:  $\ell(y \theta^\top \Phi(x)) = \mathbb{1}_{y \theta^\top \Phi(x) < 0}$
  - Usual **convex** losses:



# Main motivating examples

- **Support vector machine** (hinge loss): **non-smooth**

$$\ell(Y, \theta^\top \Phi(X)) = \max\{1 - Y\theta^\top \Phi(X), 0\}$$

- **Logistic regression**: **smooth**

$$\ell(Y, \theta^\top \Phi(X)) = \log(1 + \exp(-Y\theta^\top \Phi(X)))$$

- **Least-squares regression**

$$\ell(Y, \theta^\top \Phi(X)) = \frac{1}{2}(Y - \theta^\top \Phi(X))^2$$

- **Structured output regression**

– See Tsochantaridis et al. (2005); Lacoste-Julien et al. (2013)



# Usual regularizers

- **Main goal:** avoid overfitting
- **(squared) Euclidean norm:**  $\|\theta\|_2^2 = \sum_{j=1}^d |\theta_j|^2$ 
  - Numerically well-behaved
  - Representer theorem and kernel methods :  $\theta = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004)

# Usual regularizers

- **Main goal:** avoid overfitting
- **(squared) Euclidean norm:**  $\|\theta\|_2^2 = \sum_{j=1}^d |\theta_j|^2$ 
  - Numerically well-behaved
  - Representer theorem and kernel methods :  $\theta = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004)
- **Sparsity-inducing norms**
  - Main example:  $\ell_1$ -norm  $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$
  - Perform model selection as well as regularization
  - Non-smooth optimization and structured sparsity
  - See, e.g., Bach, Jenatton, Mairal, and Obozinski (2012b,a)

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

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

convex data fitting term + regularizer

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

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

convex data fitting term + regularizer

- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- **Two fundamental questions:** (1) computing  $\hat{\theta}$  and (2) analyzing  $\hat{\theta}$

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

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

convex data fitting term + regularizer

- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- **Two fundamental questions:** (1) computing  $\hat{\theta}$  and (2) analyzing  $\hat{\theta}$ 
  - **May be tackled simultaneously**

# Supervised machine learning

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Prediction as a linear function  $\theta^\top \Phi(x)$  of features  $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization:** find  $\hat{\theta}$  solution of

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) \text{ such that } \Omega(\theta) \leq D$$

convex data fitting term + constraint

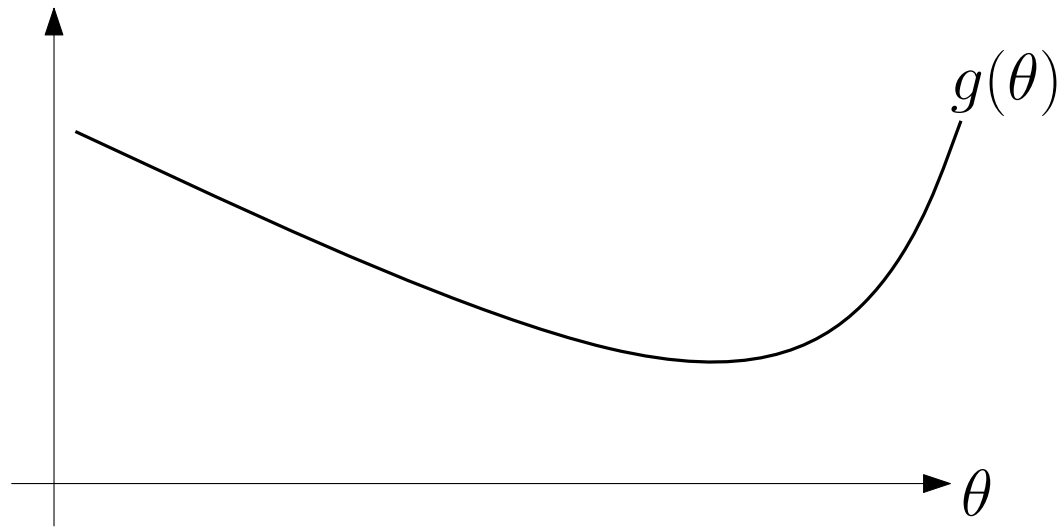
- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- **Two fundamental questions:** (1) computing  $\hat{\theta}$  and (2) analyzing  $\hat{\theta}$ 
  - **May be tackled simultaneously**

# General assumptions

- **Data:**  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ , **i.i.d.**
- Bounded features  $\Phi(x) \in \mathbb{R}^d$ :  $\|\Phi(x)\|_2 \leq R$
- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$  **training cost**
- Expected risk:  $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$  **testing cost**
- Loss for a single observation:  $f_i(\theta) = \ell(y_i, \theta^\top \Phi(x_i))$   
 $\Rightarrow \forall i, f(\theta) = \mathbb{E} f_i(\theta)$
- **Properties of  $f_i, f, \hat{f}$** 
  - **Convex** on  $\mathbb{R}^d$
  - Additional regularity assumptions: Lipschitz-continuity, smoothness and strong convexity

# Convexity

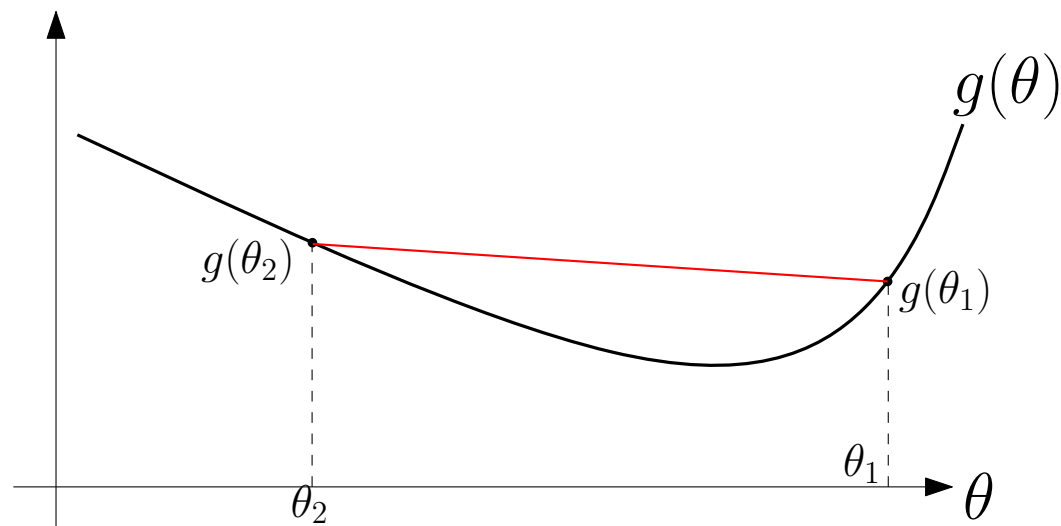
- Global definitions





# Convexity

- Global definitions (full domain)

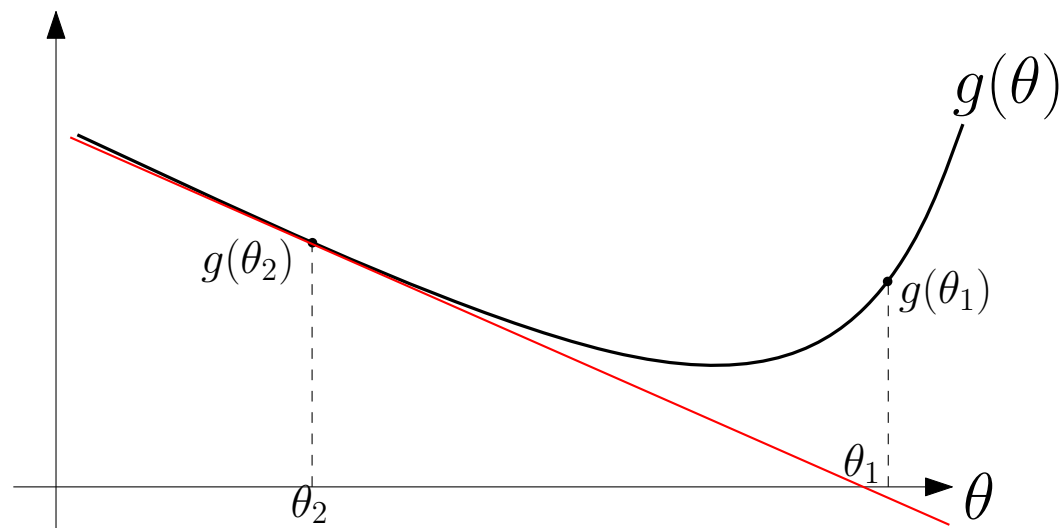


– Not assuming differentiability:

$$\forall \theta_1, \theta_2, \alpha \in [0, 1], \quad g(\alpha\theta_1 + (1 - \alpha)\theta_2) \leq \alpha g(\theta_1) + (1 - \alpha)g(\theta_2)$$

# Convexity

- **Global definitions (full domain)**



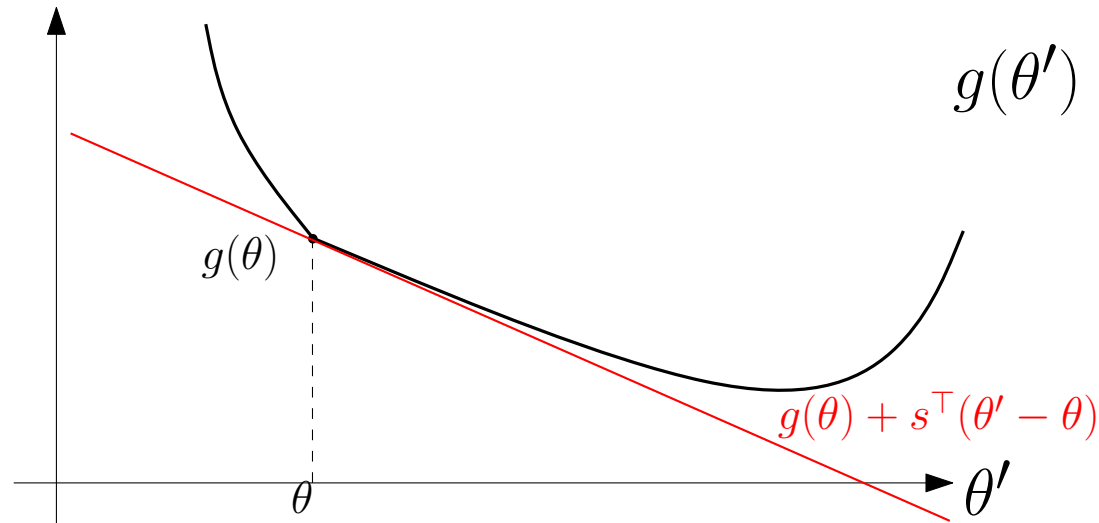
– Assuming differentiability:

$$\forall \theta_1, \theta_2, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2)$$

- **Extensions to all functions with subgradients / subdifferential**

# Subgradients and subdifferentials

- Given  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  convex



- $s \in \mathbb{R}^d$  is a **subgradient** of  $g$  at  $\theta$  if and only if

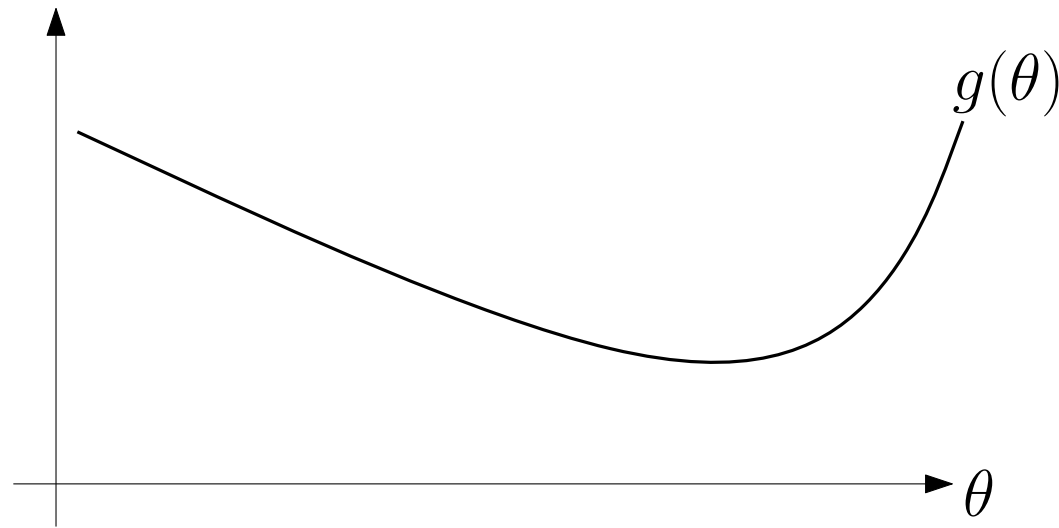
$$\forall \theta' \in \mathbb{R}^d, g(\theta') \geq g(\theta) + s^\top (\theta' - \theta)$$

- **Subdifferential**  $\partial g(\theta) =$  set of all subgradients at  $\theta$
- If  $g$  is differentiable at  $\theta$ , then  $\partial g(\theta) = \{g'(\theta)\}$
- Example: absolute value

- **The subdifferential is never empty!** See Rockafellar (1997)

# Convexity

- **Global definitions (full domain)**

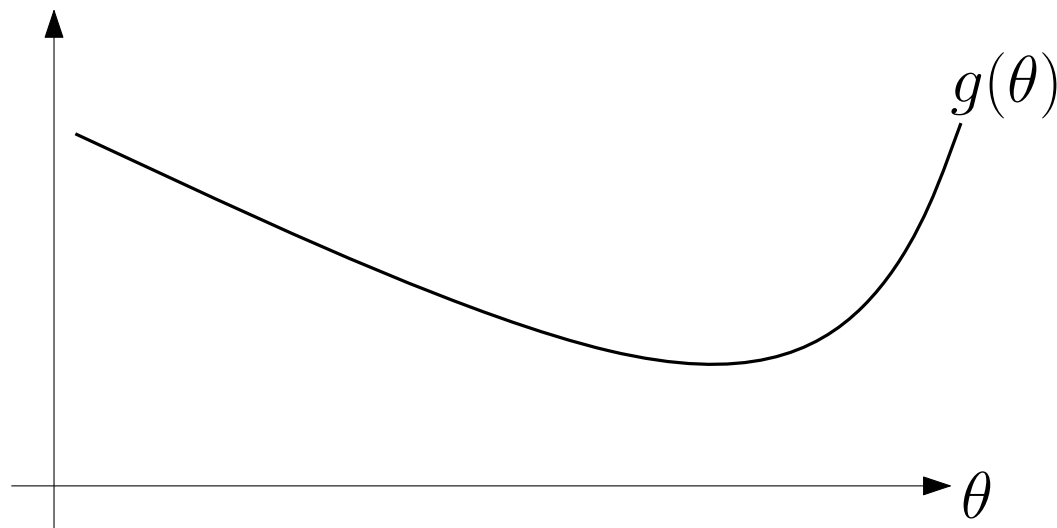


- **Local definitions**

- Twice differentiable functions
- $\forall \theta, g''(\theta) \succcurlyeq 0$  (positive semi-definite Hessians)

# Convexity

- **Global definitions (full domain)**



- **Local definitions**

- Twice differentiable functions
- $\forall \theta, g''(\theta) \succcurlyeq 0$  (positive semi-definite Hessians)

- **Why convexity?**

# Why convexity?

- **Local minimum = global minimum**
  - Optimality condition (non-smooth):  $0 \in \partial g(\theta)$
  - Optimality condition (smooth):  $g'(\theta) = 0$
- **Convex duality**
  - See Boyd and Vandenberghe (2003)
- **Recognizing convex problems**
  - See Boyd and Vandenberghe (2003)

# Lipschitz continuity

- **Bounded gradients of  $g$  ( $\Leftrightarrow$  Lipschitz-continuity):** the function  $g$  is convex, differentiable and has (sub)gradients uniformly bounded by  $B$  on the ball of center 0 and radius  $D$ :

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \Rightarrow \|g'(\theta)\|_2 \leq B$$

$\Leftrightarrow$

$$\forall \theta, \theta' \in \mathbb{R}^d, \|\theta\|_2, \|\theta'\|_2 \leq D \Rightarrow |g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|_2$$

- **Machine learning**

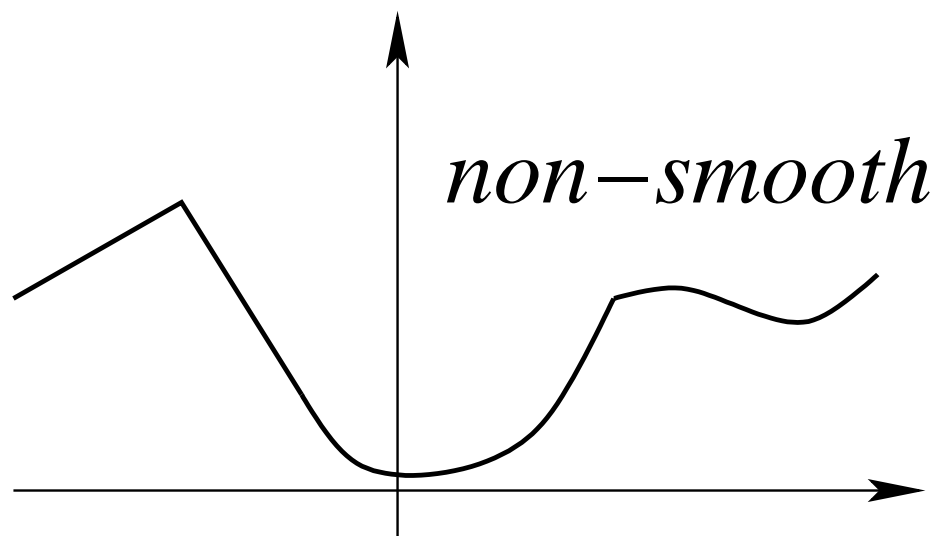
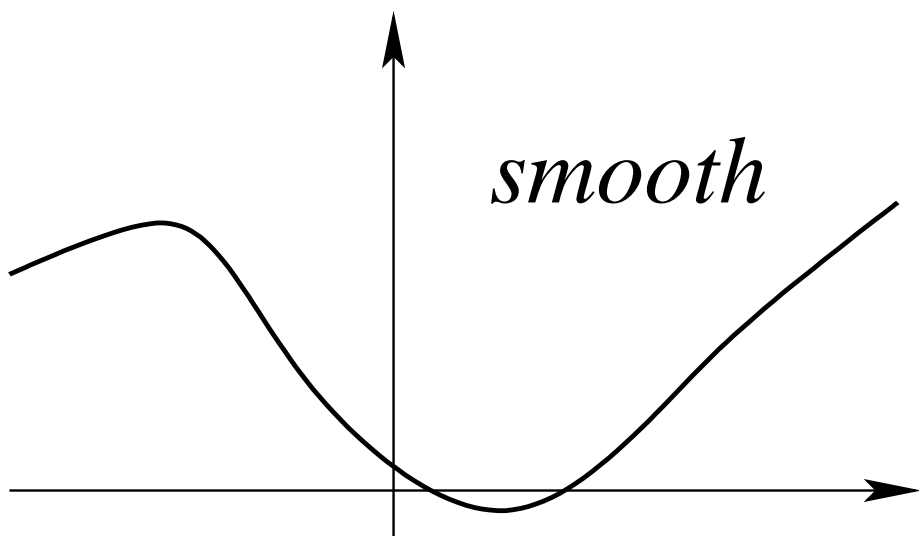
- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- $G$ -Lipschitz loss and  $R$ -bounded data:  $B = GR$

# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  **$L$ -smooth** if and only if it is differentiable and its gradient is  $L$ -Lipschitz-continuous

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L \|\theta_1 - \theta_2\|_2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, g''(\theta) \preceq L \cdot Id$





# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  **$L$ -smooth** if and only if it is differentiable and its gradient is  $L$ -Lipschitz-continuous

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L \|\theta_1 - \theta_2\|_2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \preceq L \cdot Id$

- **Machine learning**

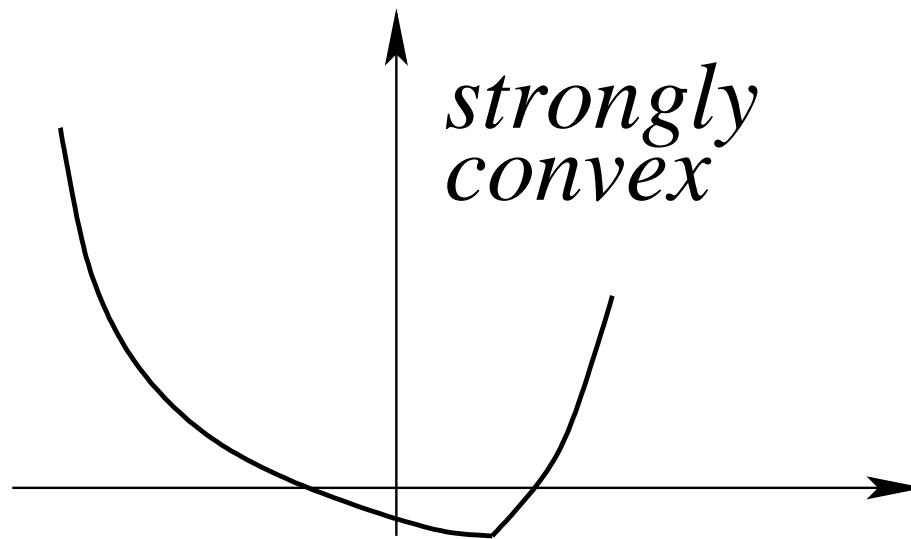
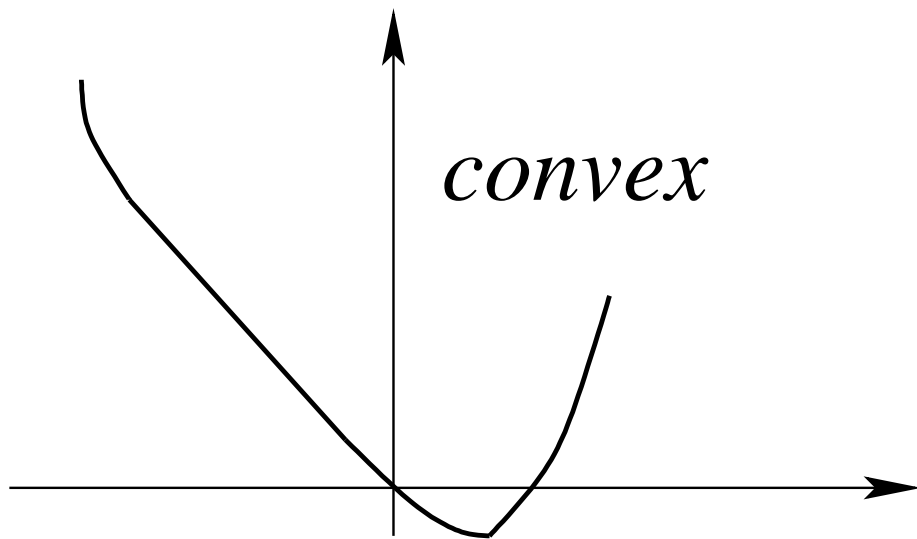
- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- Hessian  $\approx$  covariance matrix  $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$
- **$L_{\text{loss}}$ -smooth loss and  $R$ -bounded data:  $L = L_{\text{loss}} R^2$**

# Smoothness and **strong convexity**

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  **$\mu$ -strongly convex** if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, g''(\theta) \succcurlyeq \mu \cdot \text{Id}$

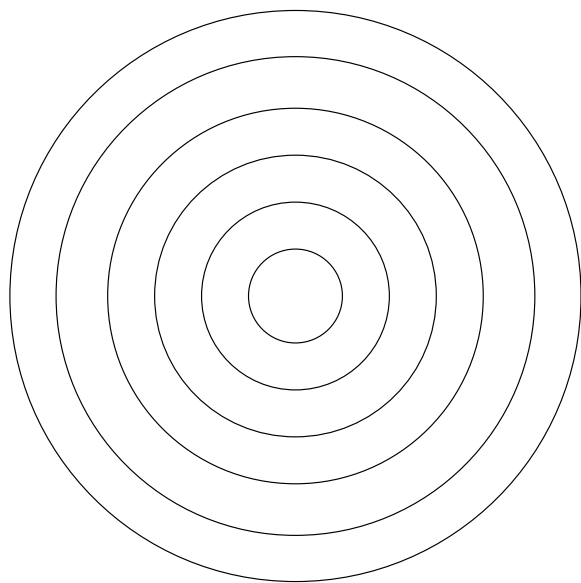


# Smoothness and strong convexity

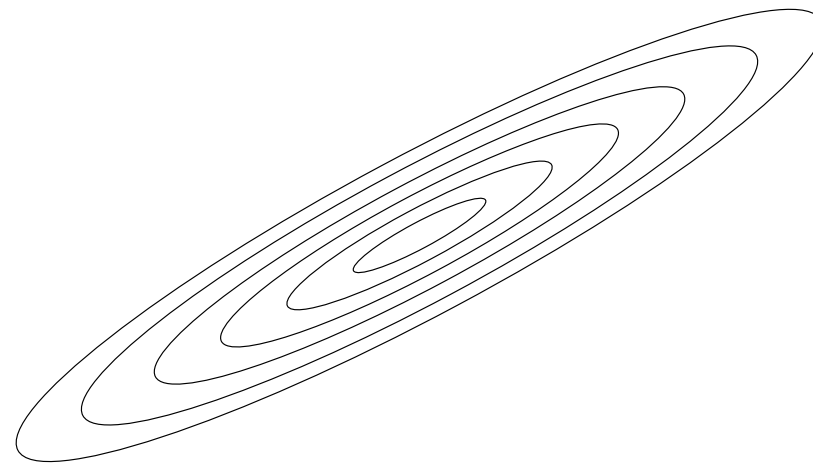
- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, g''(\theta) \succeq \mu \cdot \text{Id}$



(large  $\mu/L$ )



(small  $\mu/L$ )

# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \succeq \mu \cdot \text{Id}$

- **Machine learning**

- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- Hessian  $\approx$  covariance matrix  $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$
- **Data with invertible covariance matrix** (low correlation/dimension)

# Smoothness and strong convexity

- A function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

- If  $g$  is twice differentiable:  $\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \succeq \mu \cdot \text{Id}$

- **Machine learning**

- with  $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$

- Hessian  $\approx$  covariance matrix  $\frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$

- **Data with invertible covariance matrix** (low correlation/dimension)

- **Adding regularization by  $\frac{\mu}{2} \|\theta\|^2$**

- **creates additional bias unless  $\mu$  is small**

# Summary of smoothness/convexity assumptions

- **Bounded gradients of  $g$  (Lipschitz-continuity):** the function  $g$  is convex, differentiable and has (sub)gradients uniformly bounded by  $B$  on the ball of center  $0$  and radius  $D$ :

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \Rightarrow \|g'(\theta)\|_2 \leq B$$

- **Smoothness of  $g$ :** the function  $g$  is convex, differentiable with  $L$ -Lipschitz-continuous gradient  $g'$  (e.g., bounded Hessians):

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2$$

- **Strong convexity of  $g$ :** The function  $g$  is strongly convex with respect to the norm  $\|\cdot\|$ , with convexity constant  $\mu > 0$ :

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2$$

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

- NB: may replace  $\min_{\theta \in \mathbb{R}^d} f(\theta)$  by best (non-linear) predictions

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

**1. Uniform deviation bounds, with**  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) &= [f(\hat{\theta}) - \hat{f}(\hat{\theta})] + [\hat{f}(\hat{\theta}) - \hat{f}((\theta_*)_{\Theta})] + [\hat{f}((\theta_*)_{\Theta}) - f((\theta_*)_{\Theta})] \\ &\leq \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) + 0 + \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) \end{aligned}$$



# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

1. **Uniform deviation bounds**, with  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$

$$f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \leq \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) + \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta)$$

– Typically slow rate  $O(1/\sqrt{n})$

2. **More refined concentration results** with faster rates  $O(1/n)$

# Analysis of empirical risk minimization

- **Approximation and estimation errors:**  $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \underbrace{\left[ f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right]}_{\text{Estimation error}} + \underbrace{\left[ \min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]}_{\text{Approximation error}}$$

1. **Uniform deviation bounds**, with  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$

$$f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \leq 2 \cdot \sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)|$$

– Typically slow rate  $O(1/\sqrt{n})$

2. **More refined concentration results** with faster rates  $O(1/n)$

# Motivation from least-squares

- For least-squares, we have  $\ell(y, \theta^\top \Phi(x)) = \frac{1}{2}(y - \theta^\top \Phi(x))^2$ , and

$$\begin{aligned} f(\theta) - \hat{f}(\theta) &= \frac{1}{2} \theta^\top \left( \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top - \mathbb{E} \Phi(X) \Phi(X)^\top \right) \theta \\ &\quad - \theta^\top \left( \frac{1}{n} \sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E} Y \Phi(X) \right) + \frac{1}{2} \left( \frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E} Y^2 \right), \end{aligned}$$

$$\begin{aligned} \sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| &\leq \frac{D^2}{2} \left\| \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top - \mathbb{E} \Phi(X) \Phi(X)^\top \right\|_{\text{op}} \\ &\quad + D \left\| \frac{1}{n} \sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E} Y \Phi(X) \right\|_2 + \frac{1}{2} \left| \frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E} Y^2 \right|, \end{aligned}$$

$$\sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| \leq O(1/\sqrt{n}) \text{ with high probability from 3 concentrations}$$

# Slow rate for supervised learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - $\Omega(\theta) = \|\theta\|_2$  (Euclidean norm)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$
  - **No assumptions regarding convexity**

# Slow rate for supervised learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - $\Omega(\theta) = \|\theta\|_2$  (Euclidean norm)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$
  - **No assumptions regarding convexity**

- With probability greater than  $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{\ell_0 + GRD}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- Expected estimation error:  $\mathbb{E} \left[ \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \right] \leq \frac{4\ell_0 + 4GRD}{\sqrt{n}}$

- Using Rademacher averages (see, e.g., Boucheron et al., 2005)

- **Lipschitz functions  $\Rightarrow$  slow rate**

# Symmetrization with Rademacher variables

- Let  $\mathcal{D}' = \{x'_1, y'_1, \dots, x'_n, y'_n\}$  an **independent copy** of the data  $\mathcal{D} = \{x_1, y_1, \dots, x_n, y_n\}$ , with corresponding loss functions  $f'_i(\theta)$

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta)\right] &= \mathbb{E}\left[\sup_{\theta \in \Theta} \left(f(\theta) - \frac{1}{n} \sum_{i=1}^n f_i(\theta)\right)\right] \\ &= \mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(f'_i(\theta) - f_i(\theta) | \mathcal{D})\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n (f'_i(\theta) - f_i(\theta)) \mid \mathcal{D}\right]\right] \\ &= \mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n (f'_i(\theta) - f_i(\theta))\right] \\ &= \mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f'_i(\theta) - f_i(\theta))\right] \text{ with } \varepsilon_i \text{ uniform in } \{-1, 1\} \\ &\leq 2\mathbb{E}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta)\right] = \text{Rademacher complexity}\end{aligned}$$

# Rademacher complexity

- Rademacher complexity of the class of functions  $(X, Y) \mapsto \ell(Y, \theta^\top \Phi(X))$

$$R_n = \mathbb{E} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right]$$

- with  $f_i(\theta) = \ell(x_i, \theta^\top \Phi(x_i))$ ,  $(x_i, y_i)$ , i.i.d
- NB 1: **two** expectations, with respect to  $\mathcal{D}$  and with respect to  $\varepsilon$ 
  - “Empirical” Rademacher average  $\hat{R}_n$  by conditioning on  $\mathcal{D}$
- NB 2: sometimes defined as  $\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right|$
- **Main property:**

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) \right] \text{ and } \mathbb{E} \left[ \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) \right] \leq 2R_n$$

# From Rademacher complexity to uniform bound

- Let  $Z = \sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)|$

- By changing the pair  $(x_i, y_i)$ ,  $Z$  may only change by

$$\frac{2}{n} \sup |\ell(Y, \theta^\top \Phi(X))| \leq \frac{2}{n} (\sup |\ell(Y, 0)| + GRD) \leq \frac{2}{n} (\ell_0 + GRD) = c$$

with  $\sup |\ell(Y, 0)| = \ell_0$

- **MacDiarmid inequality:** with probability greater than  $1 - \delta$ ,

$$Z \leq \mathbb{E}Z + \sqrt{\frac{n}{2}}c \cdot \sqrt{\log \frac{1}{\delta}} \leq 2R_n + \frac{\sqrt{2}}{\sqrt{n}}(\ell_0 + GRD) \sqrt{\log \frac{1}{\delta}}$$



## Bounding the Rademacher average - I

- We have, with  $\varphi_i(u) = \ell(y_i, u) - \ell(y_i, 0)$  is almost surely  $G$ -Lipschitz:

$$\begin{aligned}\hat{R}_n &= \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right] \\ &\leq \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(0) \right] + \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f_i(\theta) - f_i(0)] \right] \\ &\leq 0 + \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f_i(\theta) - f_i(0)] \right] \\ &= 0 + \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i(\theta^\top \Phi(x_i)) \right]\end{aligned}$$

- Using Ledoux-Talagrand contraction results for Rademacher averages (since  $\varphi_i$  is  $G$ -Lipschitz), we get (Meir and Zhang, 2003):

$$\hat{R}_n \leq 2G \cdot \mathbb{E}_\varepsilon \left[ \sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \Phi(x_i) \right]$$

# Proof of Ledoux-Talagrand lemma (Meir and Zhang, 2003, Lemma 5)

- Given any  $b, a_i : \Theta \rightarrow \mathbb{R}$  (no assumption) and  $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$  any 1-Lipschitz-functions,  $i = 1, \dots, n$

$$\mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \right] \leq \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta) \right]$$

- **Proof by induction on  $n$** 
  - $n = 0$ : trivial
- From  $n$  to  $n + 1$ : see next slide

# From $n$ to $n + 1$

$$\begin{aligned}
 & \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_{n+1}} \left[ \sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n+1} \varepsilon_i \varphi_i(a_i(\theta)) \right] \\
 = & \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[ \sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))}{2} \right] \\
 = & \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[ \sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{|\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))|}{2} \right] \\
 \leq & \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[ \sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{|a_{n+1}(\theta) - a_{n+1}(\theta')|}{2} \right] \\
 = & \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \mathbb{E}_{\varepsilon_{n+1}} \left[ \sup_{\theta \in \Theta} b(\theta) + \varepsilon_{n+1} a_{n+1}(\theta) + \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \right] \\
 \leq & \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n, \varepsilon_{n+1}} \left[ \sup_{\theta \in \Theta} b(\theta) + \varepsilon_{n+1} a_{n+1}(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta) \right] \text{ by recursion}
 \end{aligned}$$

# Bounding the Rademacher average - II

- We have:

$$\begin{aligned} R_n &\leq 2G\mathbb{E}\left[\sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \Phi(x_i)\right] \\ &= 2G\mathbb{E}\left\|D\frac{1}{n} \sum_{i=1}^n \varepsilon_i \Phi(x_i)\right\|_2 \\ &\leq 2GD \sqrt{\mathbb{E}\left\|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \Phi(x_i)\right\|_2^2} \text{ by Jensen's inequality} \\ &\leq \frac{2GRD}{\sqrt{n}} \text{ by using } \|\Phi(x)\|_2 \leq R \text{ and independence} \end{aligned}$$

- Overall, we get, with probability  $1 - \delta$ :

$$\sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)| \leq \frac{1}{\sqrt{n}} (\ell_0 + GRD) \left(4 + \sqrt{2 \log \frac{1}{\delta}}\right)$$

## Putting it all together

- We have, with probability  $1 - \delta$ 
  - For exact minimizer  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$ , we have

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) &\leq \sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) + \sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) \\ &\leq \frac{2}{\sqrt{n}}(\ell_0 + GRD)(4 + \sqrt{2 \log \frac{1}{\delta}}) \end{aligned}$$

- For inexact minimizer  $\eta \in \Theta$

$$f(\eta) - \min_{\theta \in \Theta} f(\theta) \leq 2 \cdot \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| + [\hat{f}(\eta) - \hat{f}(\hat{\theta})]$$

- **Only need to optimize with precision  $\frac{2}{\sqrt{n}}(\ell_0 + GRD)$**

## Putting it all together

- We have, with probability  $1 - \delta$ 
  - For exact minimizer  $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{f}(\theta)$ , we have

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) &\leq 2 \cdot \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \\ &\leq \frac{2}{\sqrt{n}} (\ell_0 + GRD) (4 + \sqrt{2 \log \frac{1}{\delta}}) \end{aligned}$$

- For inexact minimizer  $\eta \in \Theta$

$$f(\eta) - \min_{\theta \in \Theta} f(\theta) \leq 2 \cdot \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| + [\hat{f}(\eta) - \hat{f}(\hat{\theta})]$$

- **Only need to optimize with precision  $\frac{2}{\sqrt{n}} (\ell_0 + GRD)$**

# Slow rate for supervised learning (summary)

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - $\Omega(\theta) = \|\theta\|_2$  (Euclidean norm)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$
  - **No assumptions regarding convexity**

- With probability greater than  $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{(\ell_0 + GRD)}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- Expected estimation error:  $\mathbb{E} \left[ \sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \right] \leq \frac{4(\ell_0 + GRD)}{\sqrt{n}}$

- Using Rademacher averages (see, e.g., Boucheron et al., 2005)

- **Lipschitz functions  $\Rightarrow$  slow rate**

## Motivation from mean estimation

- Estimator  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^n (\theta - z_i)^2 = \hat{f}(\theta)$
- From before:
  - $f(\theta) = \frac{1}{2} \mathbb{E}(\theta - z)^2 = \frac{1}{2}(\theta - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z) = \hat{f}(\theta) + O(1/\sqrt{n})$
  - $f(\hat{\theta}) = \frac{1}{2}(\hat{\theta} - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z) = f(\mathbb{E}z) + O(1/\sqrt{n})$



# Motivation from mean estimation

- Estimator  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i = \arg \min_{\theta \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^n (\theta - z_i)^2 = \hat{f}(\theta)$

- From before:

- $f(\theta) = \frac{1}{2} \mathbb{E}(\theta - z)^2 = \frac{1}{2}(\theta - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z) = \hat{f}(\theta) + O(1/\sqrt{n})$

- $f(\hat{\theta}) = \frac{1}{2}(\hat{\theta} - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z) = f(\mathbb{E}z) + O(1/\sqrt{n})$

- More refined/direct bound:

$$f(\hat{\theta}) - f(\mathbb{E}z) = \frac{1}{2}(\hat{\theta} - \mathbb{E}z)^2$$

$$\mathbb{E}[f(\hat{\theta}) - f(\mathbb{E}z)] = \frac{1}{2} \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n z_i - \mathbb{E}z \right)^2 = \frac{1}{2n} \text{var}(z)$$

- **Bound only at  $\hat{\theta}$  + strong convexity** (instead of uniform bound)

# Fast rate for supervised learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - Same as before (bounded features, Lipschitz loss)
  - Regularized risks:  $f^\mu(\theta) = f(\theta) + \frac{\mu}{2}\|\theta\|_2^2$  and  $\hat{f}^\mu(\theta) = \hat{f}(\theta) + \frac{\mu}{2}\|\theta\|_2^2$
  - **Convexity**
- For any  $a > 0$ , with probability greater than  $1 - \delta$ , for all  $\theta \in \mathbb{R}^d$ ,
$$f^\mu(\hat{\theta}) - \min_{\eta \in \mathbb{R}^d} f^\mu(\eta) \leq \frac{8G^2 R^2 (32 + \log \frac{1}{\delta})}{\mu n}$$
- Results from Sridharan, Srebro, and Shalev-Shwartz (2008)
  - see also Boucheron and Massart (2011) and references therein
- **Strongly convex functions  $\Rightarrow$  fast rate**
  - Warning:  $\mu$  should decrease with  $n$  to reduce approximation error

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# Complexity results in convex optimization

- **Assumption:**  $g$  convex on  $\mathbb{R}^d$
- **Classical generic algorithms**
  - Gradient descent and accelerated gradient descent
  - Newton method
  - Subgradient method and ellipsoid algorithm

# Complexity results in convex optimization

- **Assumption:**  $g$  convex on  $\mathbb{R}^d$
- **Classical generic algorithms**
  - Gradient descent and accelerated gradient descent
  - Newton method
  - Subgradient method and ellipsoid algorithm
- **Key additional properties of  $g$** 
  - Lipschitz continuity, smoothness or strong convexity
- **Key insight from Bottou and Bousquet (2008)**
  - In machine learning, no need to optimize below estimation error
- **Key references:** Nesterov (2004), Bubeck (2015)

# Several criteria for characterizing convergence

- **Objective function values**

$$g(\theta) - \inf_{\eta \in \mathbb{R}^d} g(\eta)$$

- Usually weaker condition

- **Iterates**

$$\inf_{\eta \in \arg \min g} \|\theta - \eta\|^2$$

- Typically used for strongly-convex problems

- NB: similarity with prediction vs. estimation in statistics

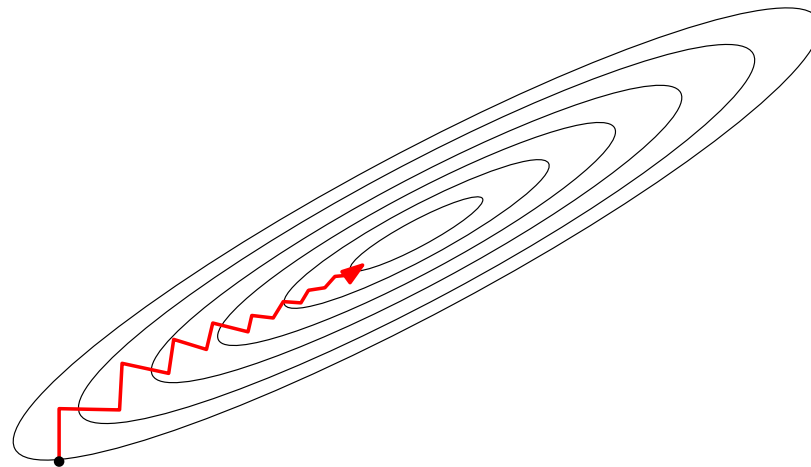
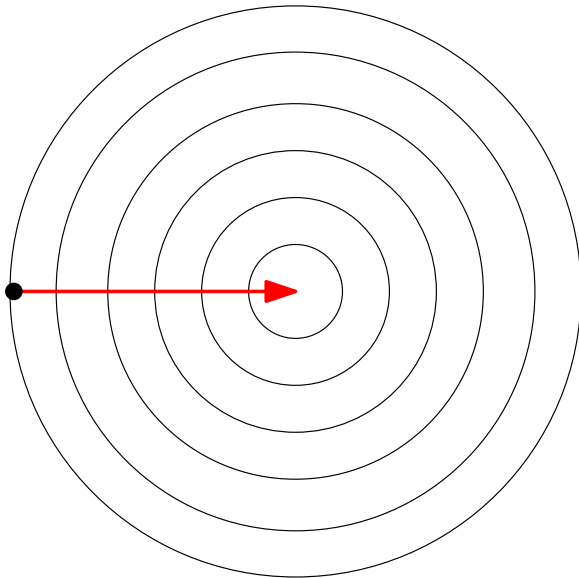
# (smooth) gradient descent

- **Assumptions**

- $g$  convex with  $L$ -Lipschitz-continuous gradient (e.g.,  $L$ -smooth)

- **Algorithm:**

$$\theta_t = \theta_{t-1} - \frac{1}{L}g'(\theta_{t-1})$$





# (smooth) gradient descent - strong convexity

- **Assumptions**

- $g$  convex with  $L$ -Lipschitz-continuous gradient (e.g.,  $L$ -smooth)
- $g$   $\mu$ -strongly convex

- **Algorithm:**

$$\theta_t = \theta_{t-1} - \frac{1}{L}g'(\theta_{t-1})$$

- **Bound:**

$$g(\theta_t) - g(\theta_*) \leq (1 - \mu/L)^t [g(\theta_0) - g(\theta_*)]$$

- Three-line proof

- **Line search, steepest descent or constant step-size**

# (smooth) gradient descent - slow rate

- **Assumptions**

- $g$  convex with  $L$ -Lipschitz-continuous gradient (e.g.,  $L$ -smooth)
- **Minimum attained at  $\theta_*$**

- **Algorithm:**

$$\theta_t = \theta_{t-1} - \frac{1}{L}g'(\theta_{t-1})$$

- **Bound:**

$$g(\theta_t) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{t + 4}$$

- Four-line proof

- **Adaptivity of gradient descent to problem difficulty**

- **Not best possible convergence rates after  $O(d)$  iterations**

# Gradient descent - Proof for quadratic functions

- Quadratic **convex** function:  $g(\theta) = \frac{1}{2}\theta^\top H\theta - c^\top \theta$ 
  - $\mu$  and  $L$  are smallest largest eigenvalues of  $H$
  - Global optimum  $\theta_* = H^{-1}c$  (or  $H^\dagger c$ )

- Gradient descent:

$$\theta_t = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - c) = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - H\theta_*)$$

$$\theta_t - \theta_* = \left(I - \frac{1}{L}H\right)(\theta_{t-1} - \theta_*) = \left(I - \frac{1}{L}H\right)^t(\theta_0 - \theta_*)$$

- **Strong convexity**  $\mu > 0$ : eigenvalues of  $\left(I - \frac{1}{L}H\right)^t$  in  $[0, (1 - \frac{\mu}{L})^t]$ 
  - Convergence of iterates:  $\|\theta_t - \theta_*\|^2 \leq (1 - \mu/L)^{2t}\|\theta_0 - \theta_*\|^2$
  - Function values:  $g(\theta_t) - g(\theta_*) \leq (1 - \mu/L)^{2t}[g(\theta_0) - g(\theta_*)]$

# Gradient descent - Proof for quadratic functions

- Quadratic **convex** function:  $g(\theta) = \frac{1}{2}\theta^\top H\theta - c^\top \theta$ 
  - $\mu$  and  $L$  are smallest largest eigenvalues of  $H$
  - Global optimum  $\theta_* = H^{-1}c$  (or  $H^\dagger c$ )

- Gradient descent:

$$\theta_t = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - c) = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - H\theta_*)$$

$$\theta_t - \theta_* = \left(I - \frac{1}{L}H\right)(\theta_{t-1} - \theta_*) = \left(I - \frac{1}{L}H\right)^t(\theta_0 - \theta_*)$$

- **Convexity**  $\mu = 0$ : eigenvalues of  $\left(I - \frac{1}{L}H\right)^t$  in  $[0, 1]$ 
  - **No convergence of iterates**:  $\|\theta_t - \theta_*\|^2 \leq \|\theta_0 - \theta_*\|^2$
  - Function values:  $g(\theta_t) - g(\theta_*) \leq \max_{v \in [0, L]} v(1 - v/L)^{2t} \|\theta_0 - \theta_*\|^2$   
 $g(\theta_t) - g(\theta_*) \leq \frac{L}{t} \|\theta_0 - \theta_*\|^2$

# Properties of smooth convex functions

- Let  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  a convex  $L$ -smooth function. Then for all  $\theta, \eta \in \mathbb{R}^d$ :
  - Definition:  $\|g'(\theta) - g'(\eta)\| \leq L\|\theta - \eta\|$
  - If twice differentiable:  $0 \preceq g''(\theta) \preceq LI$
- Quadratic upper-bound:  $0 \leq g(\theta) - g(\eta) - g'(\eta)^\top (\theta - \eta) \leq \frac{L}{2}\|\theta - \eta\|^2$ 
  - Taylor expansion with integral remainder
- Co-coercivity:  $\frac{1}{L}\|g'(\theta) - g'(\eta)\|^2 \leq [g'(\theta) - g'(\eta)]^\top (\theta - \eta)$
- If  $g$  is  $\mu$ -strongly convex, then

$$g(\theta) \leq g(\eta) + g'(\eta)^\top (\theta - \eta) + \frac{1}{2\mu}\|g'(\theta) - g'(\eta)\|^2$$

- “Distance” to optimum:  $g(\theta) - g(\theta_*) \leq g'(\theta)^\top (\theta - \theta_*)$

## Proof of co-coercivity

- Quadratic upper-bound:  $0 \leq g(\theta) - g(\eta) - g'(\eta)^\top (\theta - \eta) \leq \frac{L}{2} \|\theta - \eta\|^2$ 
  - Taylor expansion with integral remainder
- Lower bound:  $g(\theta) \geq g(\eta) + g'(\eta)^\top (\theta - \eta) + \frac{1}{2L} \|g'(\theta) - g'(\eta)\|^2$ 
  - Define  $h(\theta) = g(\theta) - \theta^\top g'(\eta)$ , convex with global minimum at  $\eta$
  - $h(\eta) \leq h(\theta - \frac{1}{L}h'(\theta)) \leq h(\theta) + h'(\theta)^\top (-\frac{1}{L}h'(\theta)) + \frac{L}{2} \|\frac{1}{L}h'(\theta)\|^2$ ,  
which is thus less than  $h(\theta) - \frac{1}{2L} \|h'(\theta)\|^2$
  - Thus  $g(\eta) - \eta^\top g'(\eta) \leq g(\theta) - \theta^\top g'(\eta) - \frac{1}{2L} \|g'(\theta) - g'(\eta)\|^2$
- Proof of co-coercivity
  - Apply lower bound twice for  $(\eta, \theta)$  and  $(\theta, \eta)$ , and sum to get  
 $0 \geq [g'(\eta) - g'(\theta)]^\top (\theta - \eta) + \frac{1}{L} \|g'(\theta) - g'(\eta)\|^2$
- NB: simple proofs with second-order derivatives

## Proof of $g(\theta) \leq g(\eta) + g'(\eta)^\top (\theta - \eta) + \frac{1}{2\mu} \|g'(\theta) - g'(\eta)\|^2$

- Define  $h(\theta) = g(\theta) - \theta^\top g'(\eta)$ , convex with global minimum at  $\eta$
- $h(\eta) = \min_{\theta} h(\theta) \geq \min_{\zeta} h(\theta) + h'(\theta)^\top (\zeta - \theta) + \frac{\mu}{2} \|\zeta - \theta\|^2$ , which is attained for  $\zeta - \theta = -\frac{1}{\mu} h'(\theta)$ 
  - This leads to  $h(\eta) \geq h(\theta) - \frac{1}{2\mu} \|h'(\theta)\|^2$
  - Hence,  $g(\eta) - \eta^\top g'(\eta) \geq g(\theta) - \theta^\top g'(\eta) - \frac{1}{2\mu} \|g'(\eta) - g'(\theta)\|^2$
  - NB: no need for smoothness
- NB: simple proofs with second-order derivatives
- With  $\eta = \theta_*$  global minimizer, another “distance” to optimum

$$g(\theta) - g(\theta_*) \leq \frac{1}{2\mu} \|g'(\theta)\|^2 \quad \text{Polyak-Lojasiewicz''}$$

# Convergence proof - gradient descent smooth strongly convex functions

- Iteration:  $\theta_t = \theta_{t-1} - \gamma g'(\theta_{t-1})$  with  $\gamma = 1/L$

$$\begin{aligned} g(\theta_t) &= g[\theta_{t-1} - \gamma g'(\theta_{t-1})] \leq g(\theta_{t-1}) + g'(\theta_{t-1})^\top [-\gamma g'(\theta_{t-1})] + \frac{L}{2} \|\gamma g'(\theta_{t-1})\|^2 \\ &= g(\theta_{t-1}) - \gamma(1 - \gamma L/2) \|g'(\theta_{t-1})\|^2 \\ &= g(\theta_{t-1}) - \frac{1}{2L} \|g'(\theta_{t-1})\|^2 \text{ if } \gamma = 1/L, \\ &\leq g(\theta_{t-1}) - \frac{\mu}{L} [g(\theta_{t-1}) - g(\theta_*)] \text{ using strongly-convex "distance" to optimum} \end{aligned}$$

Thus,  $g(\theta_t) - g(\theta_*) \leq (1 - \mu/L)^t [g(\theta_0) - g(\theta_*)]$

- May also get (Nesterov, 2004):  $\|\theta_t - \theta_*\|^2 \leq \left(1 - \frac{2\gamma\mu L}{\mu+L}\right)^t \|\theta_0 - \theta_*\|^2$   
as soon as  $\gamma \leq \frac{2}{\mu+L}$



# Convergence proof - gradient descent smooth convex functions - I

- Iteration:  $\theta_t = \theta_{t-1} - \gamma g'(\theta_{t-1})$  with  $\gamma = 1/L$

$$\begin{aligned}\|\theta_t - \theta_*\|^2 &= \|\theta_{t-1} - \theta_* - \gamma g'(\theta_{t-1})\|^2 \\ &= \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|g'(\theta_{t-1})\|^2 - 2\gamma (\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\ &\leq \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|g'(\theta_{t-1})\|^2 - 2\frac{\gamma}{L} \|g'(\theta_{t-1})\|^2 \text{ using co-coercivity} \\ &= \|\theta_{t-1} - \theta_*\|^2 - \gamma(2/L - \gamma) \|g'(\theta_{t-1})\|^2 \leq \|\theta_{t-1} - \theta_*\|^2 \text{ if } \gamma \leq 2/L \\ &\leq \|\theta_0 - \theta_*\|^2 : \text{ bounded iterates}\end{aligned}$$

$$g(\theta_t) \leq g(\theta_{t-1}) - \frac{1}{2L} \|g'(\theta_{t-1})\|^2 \text{ (see previous slide)}$$

$$g(\theta_{t-1}) - g(\theta_*) \leq g'(\theta_{t-1})^\top (\theta_{t-1} - \theta_*) \leq \|g'(\theta_{t-1})\| \cdot \|\theta_{t-1} - \theta_*\| \text{ (Cauchy-Schwarz)}$$

$$g(\theta_t) - g(\theta_*) \leq g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L \|\theta_0 - \theta_*\|^2} [g(\theta_{t-1}) - g(\theta_*)]^2$$

# Convergence proof - gradient descent smooth convex functions - II

- Iteration:  $\theta_t = \theta_{t-1} - \gamma g'(\theta_{t-1})$  with  $\gamma = 1/L$

$$g(\theta_t) - g(\theta_*) \leq g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L\|\theta_0 - \theta_*\|^2} [g(\theta_{t-1}) - g(\theta_*)]^2$$

of the form  $\Delta_k \leq \Delta_{k-1} - \alpha \Delta_{k-1}^2$  with  $0 \leq \Delta_k = g(\theta_k) - g(\theta_*) \leq \frac{L}{2} \|\theta_k - \theta_*\|^2$

$$\frac{1}{\Delta_{k-1}} \leq \frac{1}{\Delta_k} - \alpha \frac{\Delta_{k-1}}{\Delta_k} \text{ by dividing by } \Delta_k \Delta_{k-1}$$

$$\frac{1}{\Delta_{k-1}} \leq \frac{1}{\Delta_k} - \alpha \text{ because } (\Delta_k) \text{ is non-increasing}$$

$$\frac{1}{\Delta_0} \leq \frac{1}{\Delta_t} - \alpha t \text{ by summing from } k = 1 \text{ to } t$$

$$\Delta_t \leq \frac{\Delta_0}{1 + \alpha t \Delta_0} \text{ by inverting}$$

$$\leq \frac{2L\|\theta_0 - \theta_*\|^2}{t + 4} \text{ since } \Delta_0 \leq \frac{L}{2} \|\theta_0 - \theta_*\|^2 \text{ and } \alpha = \frac{1}{2L\|\theta_0 - \theta_*\|^2}$$

# Limits on convergence rate of first-order methods

- **First-order method:** any iterative algorithm that selects  $\theta_t$  in  $\theta_0 + \text{span}(f'(\theta_0), \dots, f'(\theta_{t-1}))$
- **Problem class:** convex  $L$ -smooth functions with a global minimizer  $\theta_*$
- **Theorem:** for every integer  $t \leq (d - 1)/2$  and every  $\theta_0$ , there exist functions in the problem class such that for any first-order method,

$$g(\theta_t) - g(\theta_*) \geq \frac{3}{32} \frac{L \|\theta_0 - \theta_*\|^2}{(t + 1)^2}$$

–  $O(1/t)$  rate for gradient method may not be optimal!

# Limits on convergence rate of first-order methods

## Proof sketch

- Define quadratic function

$$g_t(\theta) = \frac{L}{8} \left[ (\theta^1)^2 + \sum_{i=1}^{t-1} (\theta^i - \theta^{i+1})^2 + (\theta^t)^2 - 2\theta^1 \right]$$

- Fact 1:  $g_t$  is  $L$ -smooth
  - Fact 2: minimizer supported by first  $t$  coordinates (closed form)
  - Fact 3: any first-order method starting from zero will be supported in the first  $k$  coordinates after iteration  $k$
  - Fact 4: the minimum over this support in  $\{1, \dots, k\}$  may be computed in closed form
- Given iteration  $k$ , take  $g = g_{2k+1}$  and compute lower-bound on  $\frac{g(\theta_k) - g(\theta_*)}{\|\theta_0 - \theta_*\|^2}$

# Accelerated gradient methods (Nesterov, 1983)

- **Assumptions**

- $g$  convex with  $L$ -Lipschitz-cont. gradient , min. attained at  $\theta_*$

- **Algorithm:**

$$\theta_t = \eta_{t-1} - \frac{1}{L}g'(\eta_{t-1})$$

$$\eta_t = \theta_t + \frac{t-1}{t+2}(\theta_t - \theta_{t-1})$$

- **Bound:**

$$g(\theta_t) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{(t+1)^2}$$

- Ten-line proof (see, e.g., Schmidt, Le Roux, and Bach, 2011)

- Not improvable

- Extension to strongly-convex functions

# Accelerated gradient methods - strong convexity

- **Assumptions**

- $g$  convex with  $L$ -Lipschitz-cont. gradient , min. attained at  $\theta_*$
- $g$   $\mu$ -strongly convex

- **Algorithm:**

$$\theta_t = \eta_{t-1} - \frac{1}{L}g'(\eta_{t-1})$$
$$\eta_t = \theta_t + \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}}(\theta_t - \theta_{t-1})$$

- **Bound:**  $g(\theta_t) - f(\theta_*) \leq L\|\theta_0 - \theta_*\|^2(1 - \sqrt{\mu/L})^t$

- Ten-line proof (see, e.g., Schmidt, Le Roux, and Bach, 2011)
- Not improvable
- Relationship with conjugate gradient for quadratic functions

# Optimization for sparsity-inducing norms

(see Bach, Jenatton, Mairal, and Obozinski, 2012b)

- Gradient descent as a **proximal method** (differentiable functions)

$$- \theta_{t+1} = \arg \min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2} \|\theta - \theta_t\|_2^2$$

$$- \theta_{t+1} = \theta_t - \frac{1}{L} \nabla f(\theta_t)$$

# Optimization for sparsity-inducing norms

(see Bach, Jenatton, Mairal, and Obozinski, 2012b)

- Gradient descent as a **proximal method** (differentiable functions)

$$- \theta_{t+1} = \arg \min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2} \|\theta - \theta_t\|_2^2$$

$$- \theta_{t+1} = \theta_t - \frac{1}{L} \nabla f(\theta_t)$$

- Problems of the form:

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

$$- \theta_{t+1} = \arg \min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \mu \Omega(\theta) + \frac{L}{2} \|\theta - \theta_t\|_2^2$$

$$- \Omega(\theta) = \|\theta\|_1 \Rightarrow \text{Thresholded gradient descent}$$

- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)



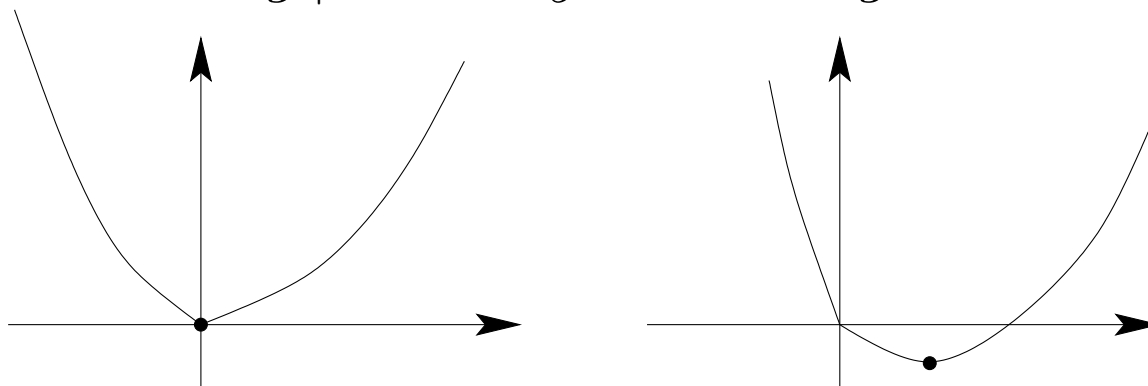
# Soft-thresholding for the $\ell_1$ -norm

- **Example 1:** quadratic problem in 1D, i.e.

$$\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|$$

- Piecewise quadratic function with a kink at zero

– Derivative at  $0+$ :  $g_+ = \lambda - y$  and  $0-$ :  $g_- = -\lambda - y$



- $x = 0$  is the solution iff  $g_+ \geq 0$  and  $g_- \leq 0$  (i.e.,  $|y| \leq \lambda$ )
- $x \geq 0$  is the solution iff  $g_+ \leq 0$  (i.e.,  $y \geq \lambda$ )  $\Rightarrow x^* = y - \lambda$
- $x \leq 0$  is the solution iff  $g_- \leq 0$  (i.e.,  $y \leq -\lambda$ )  $\Rightarrow x^* = y + \lambda$

- Solution  $x^* = \text{sign}(y)(|y| - \lambda)_+$  = soft thresholding

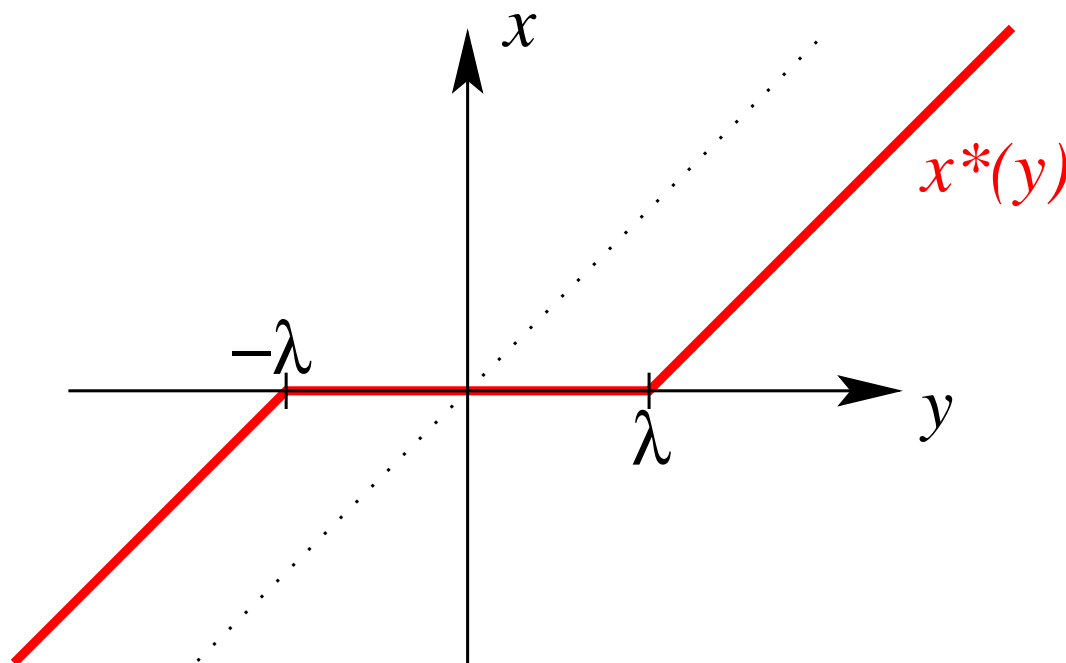
# Soft-thresholding for the $\ell_1$ -norm

- **Example 1:** quadratic problem in 1D, i.e.

$$\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|$$

- Piecewise quadratic function with a kink at zero

- Solution  $x^* = \text{sign}(y)(|y| - \lambda)_+$  = soft thresholding



# Projected gradient descent

- Problems of the form:  $\min_{\theta \in \mathcal{K}} f(\theta)$ 
  - $\theta_{t+1} = \arg \min_{\theta \in \mathcal{K}} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2} \|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \arg \min_{\theta \in \mathcal{K}} \frac{1}{2} \left\| \theta - \left( \theta_t - \frac{1}{L} \nabla f(\theta_t) \right) \right\|_2^2$
  - Projected gradient descent
- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

# Newton method

- Given  $\theta_{t-1}$ , minimize second-order Taylor expansion

$$\tilde{g}(\theta) = g(\theta_{t-1}) + g'(\theta_{t-1})^\top (\theta - \theta_{t-1}) + \frac{1}{2} (\theta - \theta_{t-1})^\top g''(\theta_{t-1}) (\theta - \theta_{t-1})$$

- **Expensive Iteration:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - Running-time complexity:  $O(d^3)$  in general
- **Quadratic convergence:** If  $\|\theta_{t-1} - \theta_*\|$  small enough, for some constant  $C$ , we have

$$(C\|\theta_t - \theta_*\|) = (C\|\theta_{t-1} - \theta_*\|)^2$$

- See Boyd and Vandenberghe (2003)

# Summary: minimizing **smooth** convex functions

- **Assumption:**  $g$  convex
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for smooth convex functions
  - $O(e^{-t\mu/L})$  convergence rate for strongly smooth convex functions
  - Optimal rates  $O(1/t^2)$  and  $O(e^{-t\sqrt{\mu/L}})$
- **Newton method:**  $\theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  convergence rate

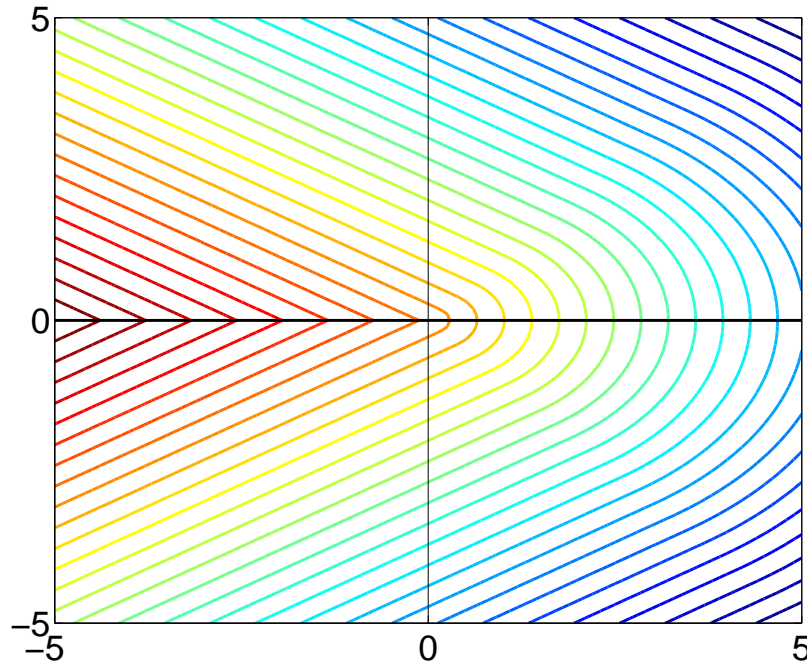
# Summary: minimizing **smooth** convex functions

- **Assumption:**  $g$  convex
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for smooth convex functions
  - $O(e^{-t\mu/L})$  convergence rate for strongly smooth convex functions
  - Optimal rates  $O(1/t^2)$  and  $O(e^{-t\sqrt{\mu/L}})$
- **Newton method:**  $\theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  convergence rate
- **From smooth to non-smooth**
  - Subgradient method and ellipsoid

# Counter-example (Bertsekas, 1999)

## Steepest descent for nonsmooth objectives

- $g(\theta_1, \theta_2) = \begin{cases} -5(9\theta_1^2 + 16\theta_2^2)^{1/2} & \text{if } \theta_1 > |\theta_2| \\ -(9\theta_1 + 16|\theta_2|)^{1/2} & \text{if } \theta_1 \leq |\theta_2| \end{cases}$
- Steepest descent starting from any  $\theta$  such that  $\theta_1 > |\theta_2| > (9/16)^2|\theta_1|$



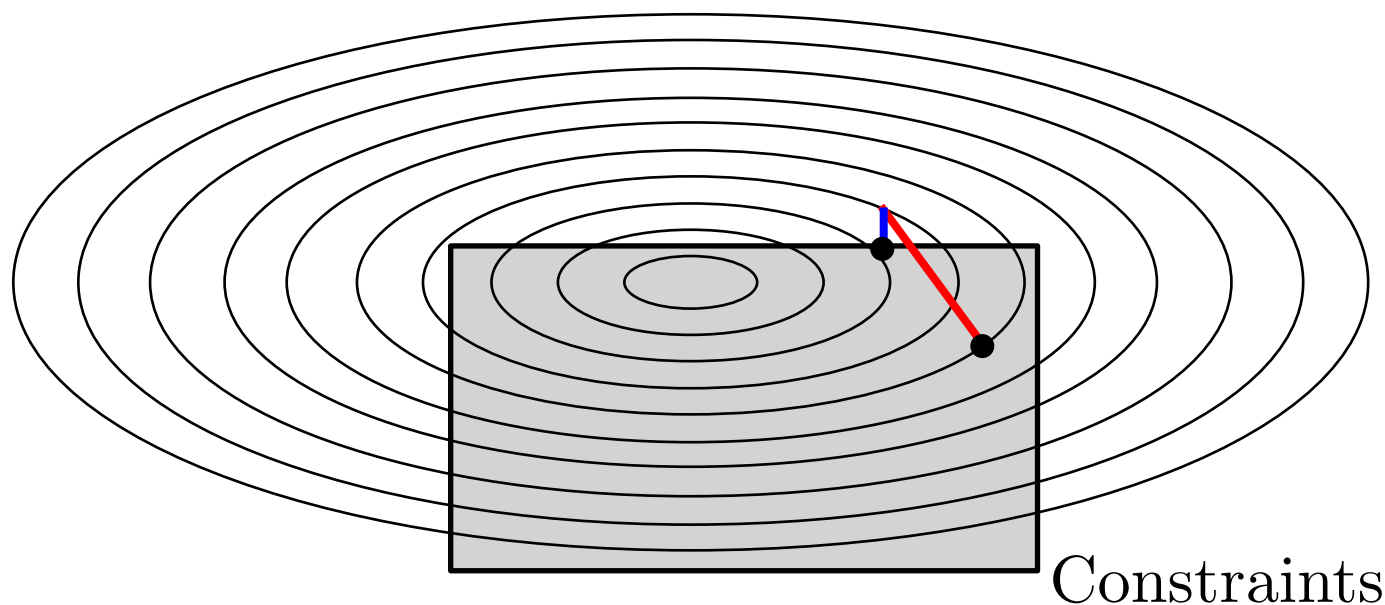
# Subgradient method/“descent” (Shor et al., 1985)

- **Assumptions**

- $g$  convex and  $B$ -Lipschitz-continuous on  $\{\|\theta\|_2 \leq D\}$

- **Algorithm:**  $\theta_t = \Pi_D \left( \theta_{t-1} - \frac{2D}{B\sqrt{t}} g'(\theta_{t-1}) \right)$

- $\Pi_D$  : orthogonal projection onto  $\{\|\theta\|_2 \leq D\}$





# Subgradient method/“descent” (Shor et al., 1985)

- **Assumptions**

- $g$  convex and  $B$ -Lipschitz-continuous on  $\{\|\theta\|_2 \leq D\}$

- **Algorithm:**  $\theta_t = \Pi_D \left( \theta_{t-1} - \frac{2D}{B\sqrt{t}} g'(\theta_{t-1}) \right)$

- $\Pi_D$  : orthogonal projection onto  $\{\|\theta\|_2 \leq D\}$

- **Bound:**

$$g \left( \frac{1}{t} \sum_{k=0}^{t-1} \theta_k \right) - g(\theta_*) \leq \frac{2DB}{\sqrt{t}}$$

- Three-line proof

- Best possible convergence rate after  $O(d)$  iterations (Bubeck, 2015)

# Subgradient method/“descent” - proof - I

- Iteration:  $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t g'(\theta_{t-1}))$  with  $\gamma_t = \frac{2D}{B\sqrt{t}}$
- Assumption:  $\|g'(\theta)\|_2 \leq B$  and  $\|\theta\|_2 \leq D$

$$\begin{aligned}\|\theta_t - \theta_*\|_2^2 &\leq \|\theta_{t-1} - \theta_* - \gamma_t g'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2 \gamma_t^2 - 2\gamma_t (\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \text{ because } \|g'(\theta_{t-1})\|_2 \leq B \\ &\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2 \gamma_t^2 - 2\gamma_t [g(\theta_{t-1}) - g(\theta_*)] \text{ (property of subgradients)}\end{aligned}$$

- leading to

$$g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2 \gamma_t}{2} + \frac{1}{2\gamma_t} [\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2]$$

# Subgradient method/“descent” - proof - II

- Starting from  $g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t} [\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2]$

- Constant step-size  $\gamma_t = \gamma$

$$\begin{aligned} \sum_{u=1}^t [g(\theta_{u-1}) - g(\theta_*)] &\leq \sum_{u=1}^t \frac{B^2\gamma}{2} + \sum_{u=1}^t \frac{1}{2\gamma} [\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2] \\ &\leq t \frac{B^2\gamma}{2} + \frac{1}{2\gamma} \|\theta_0 - \theta_*\|_2^2 \leq t \frac{B^2\gamma}{2} + \frac{2}{\gamma} D^2 \end{aligned}$$

- Optimized step-size  $\gamma_t = \frac{2D}{B\sqrt{t}}$  depends on “horizon”

– Leads to bound of  $2DB\sqrt{t}$

- Using convexity:  $g\left(\frac{1}{t} \sum_{k=0}^{t-1} \theta_k\right) - g(\theta_*) \leq \frac{2DB}{\sqrt{t}}$

# Subgradient method/“descent” - proof - III

- Starting from  $g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t} [\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2]$
- Decreasing step-size

$$\begin{aligned}
 \sum_{u=1}^t [g(\theta_{u-1}) - g(\theta_*)] &\leq \sum_{u=1}^t \frac{B^2\gamma_u}{2} + \sum_{u=1}^t \frac{1}{2\gamma_u} [\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2] \\
 &= \sum_{u=1}^t \frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1} \|\theta_u - \theta_*\|_2^2 \left( \frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u} \right) + \frac{\|\theta_0 - \theta_*\|_2^2}{2\gamma_1} - \frac{\|\theta_t - \theta_*\|_2^2}{2\gamma_t} \\
 &\leq \sum_{u=1}^t \frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1} 4D^2 \left( \frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u} \right) + \frac{4D^2}{2\gamma_1} \\
 &= \sum_{u=1}^t \frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_t} \leq 3DB\sqrt{t} \text{ with } \gamma_t = \frac{2D}{B\sqrt{t}}
 \end{aligned}$$

- Using convexity:  $g\left(\frac{1}{t} \sum_{k=0}^{t-1} \theta_k\right) - g(\theta_*) \leq \frac{3DB}{\sqrt{t}}$

# Subgradient descent for machine learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi(x_i)^\top \theta)$
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$

- **Statistics:** with probability greater than  $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{GRD}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- **Optimization:** after  $t$  iterations of subgradient method

$$\hat{f}(\hat{\theta}) - \min_{\eta \in \Theta} \hat{f}(\eta) \leq \frac{GRD}{\sqrt{t}}$$

- $t = n$  iterations, with total running-time complexity of  $O(n^2d)$

# Subgradient descent - strong convexity

- **Assumptions**

- $g$  convex and  $B$ -Lipschitz-continuous on  $\{\|\theta\|_2 \leq D\}$
- $g$   $\mu$ -strongly convex

- **Algorithm:**  $\theta_t = \Pi_D \left( \theta_{t-1} - \frac{2}{\mu(t+1)} g'(\theta_{t-1}) \right)$

- **Bound:**

$$g \left( \frac{2}{t(t+1)} \sum_{k=1}^t k \theta_{k-1} \right) - g(\theta_*) \leq \frac{2B^2}{\mu(t+1)}$$

- Three-line proof

- Best possible convergence rate after  $O(d)$  iterations (Bubeck, 2015)

# Subgradient method - strong convexity - proof - I

- Iteration:  $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t g'(\theta_{t-1}))$  with  $\gamma_t = \frac{2}{\mu(t+1)}$
- Assumption:  $\|g'(\theta)\|_2 \leq B$  and  $\|\theta\|_2 \leq D$  and  $\mu$ -strong convexity of  $f$

$$\begin{aligned}\|\theta_t - \theta_*\|_2^2 &\leq \|\theta_{t-1} - \theta_* - \gamma_t g'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2 \gamma_t^2 - 2\gamma_t (\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \text{ because } \|g'(\theta_{t-1})\|_2 \leq B \\ &\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2 \gamma_t^2 - 2\gamma_t [g(\theta_{t-1}) - g(\theta_*) + \frac{\mu}{2} \|\theta_{t-1} - \theta_*\|_2^2] \\ &\quad \text{(property of subgradients and strong convexity)}\end{aligned}$$

- leading to

$$\begin{aligned}g(\theta_{t-1}) - g(\theta_*) &\leq \frac{B^2 \gamma_t}{2} + \frac{1}{2} \left[ \frac{1}{\gamma_t} - \mu \right] \|\theta_{t-1} - \theta_*\|_2^2 - \frac{1}{2\gamma_t} \|\theta_t - \theta_*\|_2^2 \\ &\leq \frac{B^2}{\mu(t+1)} + \frac{\mu}{2} \left[ \frac{t-1}{2} \right] \|\theta_{t-1} - \theta_*\|_2^2 - \frac{\mu(t+1)}{4} \|\theta_t - \theta_*\|_2^2\end{aligned}$$

# Subgradient method - strong convexity - proof - II

- From  $g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2}{\mu(t+1)} + \frac{\mu}{2} \left[ \frac{t-1}{2} \right] \|\theta_{t-1} - \theta_*\|_2^2 - \frac{\mu(t+1)}{4} \|\theta_t - \theta_*\|_2^2$

$$\begin{aligned} \sum_{u=1}^t u [g(\theta_{u-1}) - g(\theta_*)] &\leq \sum_{t=1}^u \frac{B^2 u}{\mu(u+1)} + \frac{1}{4} \sum_{u=1}^t [u(u-1) \|\theta_{u-1} - \theta_*\|_2^2 - u(u+1) \|\theta_u - \theta_*\|_2^2] \\ &\leq \frac{B^2 t}{\mu} + \frac{1}{4} [0 - t(t+1) \|\theta_t - \theta_*\|_2^2] \leq \frac{B^2 t}{\mu} \end{aligned}$$

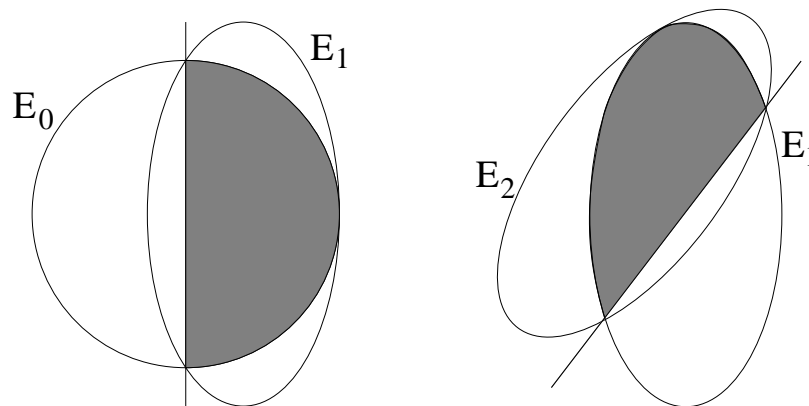
- Using convexity:  $g\left(\frac{2}{t(t+1)} \sum_{u=1}^t u \theta_{u-1}\right) - g(\theta_*) \leq \frac{2B^2}{t+1}$

- NB: with step-size  $\gamma_n = 1/(n\mu)$ , extra logarithmic factor



# Ellipsoid method

- Minimizing convex function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ 
  - Builds a sequence of ellipsoids that contains the global minima.



- Represent  $E_t = \{\theta \in \mathbb{R}^d, (\theta - \theta_t)^\top P_t^{-1}(\theta - \theta_t) \leq 1\}$
- Fact 1:  $\theta_{t+1} = \theta_t - \frac{1}{d+1}P_t h_t$  and  $P_{t+1} = \frac{d^2}{d^2-1}(P_t - \frac{2}{d+1}P_t h_t h_t^\top P_t)$   
with  $h_t = \frac{1}{\sqrt{g'(\theta_t)^\top P_t g'(\theta_t)}}g'(\theta_t)$
- Fact 2:  $\text{vol}(\mathcal{E}_t) \approx \text{vol}(\mathcal{E}_{t-1})e^{-1/2d} \Rightarrow$  CV rate in  $O(e^{-t/d^2})$

# Summary: minimizing convex functions

- **Assumption:**  $g$  convex
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/\sqrt{t})$  convergence rate for non-smooth convex functions
  - $O(1/t)$  convergence rate for smooth convex functions
  - $O(e^{-\rho t})$  convergence rate for strongly smooth convex functions
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  convergence rate

# Summary: minimizing convex functions

- **Assumption:**  $g$  convex
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/\sqrt{t})$  convergence rate for non-smooth convex functions
  - $O(1/t)$  convergence rate for smooth convex functions
  - $O(e^{-\rho t})$  convergence rate for strongly smooth convex functions
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  convergence rate
- **Key insights from Bottou and Bousquet (2008)**
  1. In machine learning, no need to optimize below statistical error
  2. In machine learning, cost functions are averages

$\Rightarrow$  **Stochastic approximation**

# Summary of rates of convergence

- Problem parameters
  - $D$  diameter of the domain
  - $B$  Lipschitz-constant
  - $L$  smoothness constant
  - $\mu$  strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$	deterministic: $B^2/(t\mu)$
smooth	deterministic: $LD^2/t^2$	deterministic: $\exp(-t\sqrt{\mu/L})$
quadratic	deterministic: $LD^2/t^2$	deterministic: $\exp(-t\sqrt{\mu/L})$

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# Stochastic approximation

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

# Stochastic approximation

- **Goal:** Minimizing a function  $f$  defined on  $\mathbb{R}^d$ 
  - given only unbiased estimates  $f'_n(\theta_n)$  of its gradients  $f'(\theta_n)$  at certain points  $\theta_n \in \mathbb{R}^d$
- **Machine learning - statistics**
  - **loss for a single pair of observations:**  $f_n(\theta) = \ell(y_n, \theta^\top \Phi(x_n))$
  - $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E} \ell(y_n, \theta^\top \Phi(x_n)) =$  **generalization error**
  - Expected gradient:  $f'(\theta) = \mathbb{E} f'_n(\theta) = \mathbb{E} \{ \ell'(y_n, \theta^\top \Phi(x_n)) \Phi(x_n) \}$
  - Non-asymptotic results
- **Number of iterations = number of observations**



# Stochastic approximation

- **Goal:** Minimizing a function  $f$  defined on  $\mathbb{R}^d$ 
  - given only unbiased estimates  $f'_n(\theta_n)$  of its gradients  $f'(\theta_n)$  at certain points  $\theta_n \in \mathbb{R}^d$
- **Stochastic approximation**
  - (much) broader applicability beyond convex optimization
    - $$\theta_n = \theta_{n-1} - \gamma_n h_n(\theta_{n-1}) \text{ with } \mathbb{E}[h_n(\theta_{n-1}) | \theta_{n-1}] = h(\theta_{n-1})$$
  - Beyond convex problems, i.i.d assumption, finite dimension, etc.
  - Typically asymptotic results (see next lecture)
  - See, e.g., Kushner and Yin (2003); Benveniste et al. (2012)

# Relationship to online learning

- **Stochastic approximation**

- Minimize  $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$  **generalization error** of  $\theta$
- Using the gradients of single i.i.d. observations

# Relationship to online learning

- **Stochastic approximation**

- Minimize  $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$  **generalization error** of  $\theta$
- Using the gradients of single i.i.d. observations

- **Batch learning**

- Finite set of observations:  $z_1, \dots, z_n$
- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^n \ell(\theta, z_i)$
- Estimator  $\hat{\theta} =$  Minimizer of  $\hat{f}(\theta)$  over a certain class  $\Theta$
- Generalization bound using uniform concentration results

# Relationship to online learning

- **Stochastic approximation**

- Minimize  $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$  **generalization error** of  $\theta$
- Using the gradients of single i.i.d. observations

- **Batch learning**

- Finite set of observations:  $z_1, \dots, z_n$
- Empirical risk:  $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^n \ell(\theta, z_k)$
- Estimator  $\hat{\theta} =$  Minimizer of  $\hat{f}(\theta)$  over a certain class  $\Theta$
- Generalization bound using uniform concentration results

- **Online learning**

- Update  $\hat{\theta}_n$  after each new (**potentially adversarial**) observation  $z_n$
- Cumulative loss:  $\frac{1}{n} \sum_{k=1}^n \ell(\hat{\theta}_{k-1}, z_k)$
- Online to batch through averaging (Cesa-Bianchi et al., 2004)

# Convex stochastic approximation

- Key properties of  $f$  and/or  $f_n$ 
  - Smoothness:  $f$   $B$ -Lipschitz continuous,  $f'$   $L$ -Lipschitz continuous
  - Strong convexity:  $f$   $\mu$ -strongly convex

# Convex stochastic approximation

- **Key properties of  $f$  and/or  $f_n$** 
  - **Smoothness**:  $f$   $B$ -Lipschitz continuous,  $f'$   $L$ -Lipschitz continuous
  - **Strong convexity**:  $f$   $\mu$ -strongly convex
- **Key algorithm**: Stochastic gradient descent (a.k.a. Robbins-Monro)

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

– Polyak-Ruppert averaging:  $\bar{\theta}_n = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k$

– Which learning rate sequence  $\gamma_n$ ? Classical setting:

$$\gamma_n = C n^{-\alpha}$$

# Convex stochastic approximation

- **Key properties of  $f$  and/or  $f_n$** 
  - **Smoothness**:  $f$   $B$ -Lipschitz continuous,  $f'$   $L$ -Lipschitz continuous
  - **Strong convexity**:  $f$   $\mu$ -strongly convex
- **Key algorithm**: Stochastic gradient descent (a.k.a. Robbins-Monro)

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

- Polyak-Ruppert averaging:  $\bar{\theta}_n = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k$
  - Which learning rate sequence  $\gamma_n$ ? Classical setting:  $\gamma_n = Cn^{-\alpha}$
- **Desirable practical behavior**
    - Applicable (at least) to classical supervised learning problems
    - Robustness to (potentially unknown) constants  $(L, B, \mu)$
    - Adaptivity to difficulty of the problem (e.g., strong convexity)

# Stochastic subgradient “descent” / method

- **Assumptions**

- $f_n$  convex and  $B$ -Lipschitz-continuous on  $\{\|\theta\|_2 \leq D\}$
- $(f_n)$  i.i.d. functions such that  $\mathbb{E}f_n = f$
- $\theta_*$  global optimum of  $f$  on  $\mathcal{C} = \{\|\theta\|_2 \leq D\}$

- **Algorithm:**  $\theta_n = \Pi_D \left( \theta_{n-1} - \frac{2D}{B\sqrt{n}} f'_n(\theta_{n-1}) \right)$



# Stochastic subgradient “descent” / method

- **Assumptions**

- $f_n$  convex and  $B$ -Lipschitz-continuous on  $\{\|\theta\|_2 \leq D\}$
- $(f_n)$  i.i.d. functions such that  $\mathbb{E}f_n = f$
- $\theta_*$  global optimum of  $f$  on  $\mathcal{C} = \{\|\theta\|_2 \leq D\}$

- **Algorithm:**  $\theta_n = \Pi_D \left( \theta_{n-1} - \frac{2D}{B\sqrt{n}} f'_n(\theta_{n-1}) \right)$

- **Bound:**

$$\mathbb{E}f \left( \frac{1}{n} \sum_{k=0}^{n-1} \theta_k \right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$$

- “Same” three-line proof as in the deterministic case
- **Minimax rate** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
- Running-time complexity:  $O(dn)$  after  $n$  iterations

# Stochastic subgradient method - proof - I

- Iteration:  $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f'_n(\theta_{n-1}))$  with  $\gamma_n = \frac{2D}{B\sqrt{n}}$
- $\mathcal{F}_n$  : information up to time  $n$
- $\|f'_n(\theta)\|_2 \leq B$  and  $\|\theta\|_2 \leq D$ , unbiased gradients/functions  $\mathbb{E}(f_n | \mathcal{F}_{n-1}) = f$

$$\begin{aligned} \|\theta_n - \theta_*\|_2^2 &\leq \|\theta_{n-1} - \theta_* - \gamma_n f'_n(\theta_{n-1})\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \text{ because } \|f'_n(\theta_{n-1})\|_2 \leq B \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\|\theta_n - \theta_*\|_2^2 | \mathcal{F}_{n-1}] &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'(\theta_{n-1}) \\ &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n [f(\theta_{n-1}) - f(\theta_*)] \text{ (subgradient property)} \end{aligned}$$

$$\mathbb{E}\|\theta_n - \theta_*\|_2^2 \leq \mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n [\mathbb{E}f(\theta_{n-1}) - f(\theta_*)]$$

- leading to  $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leq \frac{B^2 \gamma_n}{2} + \frac{1}{2\gamma_n} [\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_n - \theta_*\|_2^2]$

# Stochastic subgradient method - proof - II

- Starting from  $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n} [\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_n - \theta_*\|_2^2]$

$$\begin{aligned} \sum_{u=1}^n [\mathbb{E}f(\theta_{u-1}) - f(\theta_*)] &\leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \sum_{u=1}^n \frac{1}{2\gamma_u} [\mathbb{E}\|\theta_{u-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_u - \theta_*\|_2^2] \\ &\leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} \leq 2DB\sqrt{n} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}} \end{aligned}$$

- Using convexity:  $\mathbb{E}f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$

# Stochastic subgradient method

## Extension to online learning

- Assume **different and arbitrary** functions  $f_n : \mathbb{R}^d \rightarrow \mathbb{R}$ 
  - Observations of  $f'_n(\theta_{n-1}) + \varepsilon_n$
  - with  $\mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0$  and  $\|f'_n(\theta_{n-1}) + \varepsilon_n\| \leq B$  almost surely
- **Performance criterion: (normalized) regret**

$$\frac{1}{n} \sum_{i=1}^n f_i(\theta_{i-1}) - \inf_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_{i=1}^n f_i(\theta)$$

- Warning: often not normalized
- May not be non-negative (typically is)

# Stochastic subgradient method - online learning - I

- Iteration:  $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n(f'_n(\theta_{n-1}) + \varepsilon_n))$  with  $\gamma_n = \frac{2D}{B\sqrt{n}}$
- $\mathcal{F}_n$  : information up to time  $n$  -  $\theta$  an **arbitrary** point such that  $\|\theta\| \leq D$
- $\|f'_n(\theta_{n-1}) + \varepsilon_n\|_2 \leq B$  and  $\|\theta\|_2 \leq D$ , unbiased gradients  $\mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0$

$$\begin{aligned} \|\theta_n - \theta\|_2^2 &\leq \|\theta_{n-1} - \theta - \gamma_n(f'_n(\theta_{n-1}) + \varepsilon_n)\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta)^\top (f'_n(\theta_{n-1}) + \varepsilon_n) \text{ because } \|f'_n(\theta_{n-1}) + \varepsilon_n\|_2 \leq B \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\|\theta_n - \theta\|_2^2 | \mathcal{F}_{n-1}] &\leq \|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta)^\top f'_n(\theta_{n-1}) \\ &\leq \|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f_n(\theta_{n-1}) - f_n(\theta)] \text{ (subgradient property)} \end{aligned}$$

$$\mathbb{E}\|\theta_n - \theta\|_2^2 \leq \mathbb{E}\|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta)]$$

- leading to  $\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}[\mathbb{E}\|\theta_{n-1} - \theta\|_2^2 - \mathbb{E}\|\theta_n - \theta\|_2^2]$

# Stochastic subgradient method - online learning - II

- Starting from  $\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n} [\mathbb{E}\|\theta_{n-1} - \theta\|_2^2 - \mathbb{E}\|\theta_n - \theta\|_2^2]$

$$\begin{aligned} \sum_{u=1}^n [\mathbb{E}f_u(\theta_{u-1}) - f_u(\theta)] &\leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \sum_{u=1}^n \frac{1}{2\gamma_u} [\mathbb{E}\|\theta_{u-1} - \theta\|_2^2 - \mathbb{E}\|\theta_u - \theta\|_2^2] \\ &\leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} \leq 2DB\sqrt{n} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}} \end{aligned}$$

- For any  $\theta$  such that  $\|\theta\| \leq D$ :  $\frac{1}{n} \sum_{k=1}^n \mathbb{E}f_k(\theta_{k-1}) - \frac{1}{n} \sum_{k=1}^n f_k(\theta) \leq \frac{2DB}{\sqrt{n}}$

- Online to batch conversion: assuming convexity

# Stochastic subgradient descent - strong convexity - I

- **Assumptions**

- $f_n$  convex and  $B$ -Lipschitz-continuous
- $(f_n)$  i.i.d. functions such that  $\mathbb{E}f_n = f$
- $f$   $\mu$ -strongly convex on  $\{\|\theta\|_2 \leq D\}$
- $\theta_*$  global optimum of  $f$  over  $\{\|\theta\|_2 \leq D\}$

- **Algorithm:**  $\theta_n = \Pi_D \left( \theta_{n-1} - \frac{2}{\mu(n+1)} f'_n(\theta_{n-1}) \right)$

- **Bound:**

$$\mathbb{E}f \left( \frac{2}{n(n+1)} \sum_{k=1}^n k\theta_{k-1} \right) - f(\theta_*) \leq \frac{2B^2}{\mu(n+1)}$$

- “Same” proof than deterministic case (Lacoste-Julien et al., 2012)
- **Minimax rate** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)

# Stochastic subgradient - strong convexity - proof - I

- Iteration:  $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f'_n(\theta_{n-1}))$  with  $\gamma_n = \frac{2}{\mu(n+1)}$

- Assumption:  $\|f'_n(\theta)\|_2 \leq B$  and  $\|\theta\|_2 \leq D$  and  $\mu$ -strong convexity of  $f$

$$\begin{aligned} \|\theta_n - \theta_*\|_2^2 &\leq \|\theta_{n-1} - \theta_* - \gamma_n f'_n(\theta_{n-1})\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \text{ because } \|f'_n(\theta_{n-1})\|_2 \leq B \\ \mathbb{E}(\cdot | \mathcal{F}_{n-1}) &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n [f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2} \|\theta_{n-1} - \theta_*\|_2^2] \\ &\quad \text{(property of subgradients and strong convexity)} \end{aligned}$$

- leading to

$$\begin{aligned} \mathbb{E}f(\theta_{n-1}) - f(\theta_*) &\leq \frac{B^2 \gamma_n}{2} + \frac{1}{2} \left[ \frac{1}{\gamma_n} - \mu \right] \|\theta_{n-1} - \theta_*\|_2^2 - \frac{1}{2\gamma_n} \|\theta_n - \theta_*\|_2^2 \\ &\leq \frac{B^2}{\mu(n+1)} + \frac{\mu}{2} \left[ \frac{n-1}{2} \right] \|\theta_{n-1} - \theta_*\|_2^2 - \frac{\mu(n+1)}{4} \|\theta_n - \theta_*\|_2^2 \end{aligned}$$



# Stochastic subgradient - strong convexity - proof - II

- From  $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leq \frac{B^2}{\mu(n+1)} + \frac{\mu}{2} \left[ \frac{n-1}{2} \right] \mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \frac{\mu(n+1)}{4} \mathbb{E}\|\theta_n - \theta_*\|_2^2$

$$\begin{aligned} \sum_{u=1}^n u [\mathbb{E}f(\theta_{u-1}) - f(\theta_*)] &\leq \sum_{u=1}^n \frac{B^2 u}{\mu(u+1)} + \frac{1}{4} \sum_{u=1}^n [u(u-1) \mathbb{E}\|\theta_{u-1} - \theta_*\|_2^2 - u(u+1) \mathbb{E}\|\theta_u - \theta_*\|_2^2] \\ &\leq \frac{B^2 n}{\mu} + \frac{1}{4} [0 - n(n+1) \mathbb{E}\|\theta_n - \theta_*\|_2^2] \leq \frac{B^2 n}{\mu} \end{aligned}$$

- Using convexity:  $\mathbb{E}f\left(\frac{2}{n(n+1)} \sum_{u=1}^n u \theta_{u-1}\right) - g(\theta_*) \leq \frac{2B^2}{n+1}$

- NB: with step-size  $\gamma_n = 1/(n\mu)$ , extra logarithmic factor (see later)

# Stochastic subgradient descent - strong convexity - II

- **Assumptions**

- $f_n$  convex and  $B$ -Lipschitz-continuous
- $(f_n)$  i.i.d. functions such that  $\mathbb{E}f_n = f$
- $\theta_*$  global optimum of  $g = f + \frac{\mu}{2}\|\cdot\|_2^2$
- No compactness assumption - no projections

- **Algorithm:**

$$\theta_n = \theta_{n-1} - \frac{2}{\mu(n+1)} g'_n(\theta_{n-1}) = \theta_{n-1} - \frac{2}{\mu(n+1)} [f'_n(\theta_{n-1}) + \mu\theta_{n-1}]$$

- **Bound:**  $\mathbb{E}g\left(\frac{2}{n(n+1)} \sum_{k=1}^n k\theta_{k-1}\right) - g(\theta_*) \leq \frac{2B^2}{\mu(n+1)}$

- **Minimax convergence rate**

# Strong convexity - proof with $\log n$ factor - I

- Iteration:  $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f'_n(\theta_{n-1}))$  with  $\gamma_n = \frac{1}{\mu n}$
- Assumption:  $\|f'_n(\theta)\|_2 \leq B$  and  $\|\theta\|_2 \leq D$  and  $\mu$ -strong convexity of  $f$

$$\begin{aligned} \|\theta_n - \theta_*\|_2^2 &\leq \|\theta_{n-1} - \theta_* - \gamma_n f'_n(\theta_{n-1})\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \text{ because } \|f'_n(\theta_{n-1})\|_2 \leq B \\ \mathbb{E}(\cdot | \mathcal{F}_{n-1}) &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n [f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2} \|\theta_{n-1} - \theta_*\|_2^2] \\ &\quad \text{(property of subgradients and strong convexity)} \end{aligned}$$

- leading to

$$\begin{aligned} \mathbb{E}f(\theta_{n-1}) - f(\theta_*) &\leq \frac{B^2 \gamma_n}{2} + \frac{1}{2} \left[ \frac{1}{\gamma_n} - \mu \right] \|\theta_{n-1} - \theta_*\|_2^2 - \frac{1}{2\gamma_n} \|\theta_n - \theta_*\|_2^2 \\ &\leq \frac{B^2}{2\mu n} + \frac{\mu}{2} [n - 1] \|\theta_{n-1} - \theta_*\|_2^2 - \frac{n\mu}{2} \|\theta_n - \theta_*\|_2^2 \end{aligned}$$

# Strong convexity - proof with $\log n$ factor - II

- From  $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leq \frac{B^2}{2\mu n} + \frac{\mu}{2}[n-1]\|\theta_{n-1} - \theta_*\|_2^2 - \frac{n\mu}{2}\|\theta_n - \theta_*\|_2^2$

$$\begin{aligned} \sum_{u=1}^n [\mathbb{E}f(\theta_{u-1}) - f(\theta_*)] &\leq \sum_{u=1}^n \frac{B^2}{2\mu u} + \frac{1}{2} \sum_{u=1}^n [(u-1)\mathbb{E}\|\theta_{u-1} - \theta_*\|_2^2 - u\mathbb{E}\|\theta_u - \theta_*\|_2^2] \\ &\leq \frac{B^2 \log n}{2\mu} + \frac{1}{2} [0 - n\mathbb{E}\|\theta_n - \theta_*\|_2^2] \leq \frac{B^2 \log n}{2\mu} \end{aligned}$$

- Using convexity:  $\mathbb{E}f\left(\frac{1}{n} \sum_{u=1}^n \theta_{u-1}\right) - f(\theta_*) \leq \frac{B^2 \log n}{2\mu n}$

- Why could this be useful?

# Stochastic subgradient descent - strong convexity

## Online learning

- Need  $\log n$  term for uniform averaging. For all  $\theta$ :

$$\frac{1}{n} \sum_{i=1}^n f_i(\theta_{i-1}) - \frac{1}{n} \sum_{i=1}^n f_i(\theta) \leq \frac{B^2 \log n}{2\mu n}$$

- Optimal. See Hazan and Kale (2014).

# Beyond convergence in expectation

- **Typical result:**  $\mathbb{E} f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$

- Obtained with simple conditioning arguments

- **High-probability bounds**

- Markov inequality:  $\mathbb{P}\left(f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k\right) - f(\theta_*) \geq \varepsilon\right) \leq \frac{2DB}{\sqrt{n}\varepsilon}$

# Beyond convergence in expectation

- **Typical result:**  $\mathbb{E} f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$

- Obtained with simple conditioning arguments

- **High-probability bounds**

- Markov inequality:  $\mathbb{P}\left(f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k\right) - f(\theta_*) \geq \varepsilon\right) \leq \frac{2DB}{\sqrt{n}\varepsilon}$

- Deviation inequality (Nemirovski et al., 2009; Nesterov and Vial, 2008)

$$\mathbb{P}\left(f\left(\frac{1}{n} \sum_{k=0}^{n-1} \theta_k\right) - f(\theta_*) \geq \frac{2DB}{\sqrt{n}}(2 + 4t)\right) \leq 2 \exp(-t^2)$$

- See also Bach (2013) for logistic regression

# Stochastic subgradient method - high probability - I

- Iteration:  $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f'_n(\theta_{n-1}))$  with  $\gamma_n = \frac{2D}{B\sqrt{n}}$
- $\mathcal{F}_n$  : information up to time  $n$
- $\|f'_n(\theta)\|_2 \leq B$  and  $\|\theta\|_2 \leq D$ , unbiased gradients/functions  $\mathbb{E}(f_n|\mathcal{F}_{n-1}) = f$

$$\begin{aligned}\|\theta_n - \theta_*\|_2^2 &\leq \|\theta_{n-1} - \theta_* - \gamma_n f'_n(\theta_{n-1})\|_2^2 \text{ by contractivity of projections} \\ &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \text{ because } \|f'_n(\theta_{n-1})\|_2 \leq B\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\|\theta_n - \theta_*\|_2^2 | \mathcal{F}_{n-1}] &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f'(\theta_{n-1}) \\ &\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f(\theta_{n-1}) - f(\theta_*)] \text{ (subgradient property)}\end{aligned}$$

- **Without expectations** and with  $Z_n = -2\gamma_n(\theta_{n-1} - \theta_*)^\top [f'_n(\theta_{n-1}) - f'(\theta_{n-1})]$

$$\|\theta_n - \theta_*\|_2^2 \leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f(\theta_{n-1}) - f(\theta_*)] + Z_n$$



# Stochastic subgradient method - high probability - II

- Without expectations and with  $Z_n = -2\gamma_n(\theta_{n-1} - \theta_*)^\top [f'_n(\theta_{n-1}) - f'(\theta_{n-1})]$

$$\|\theta_n - \theta_*\|_2^2 \leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f(\theta_{n-1}) - f(\theta_*)] + Z_n$$

$$f(\theta_{n-1}) - f(\theta_*) \leq \frac{1}{2\gamma_n} [\|\theta_{n-1} - \theta_*\|_2^2 - \|\theta_n - \theta_*\|_2^2] + \frac{B^2\gamma_n}{2} + \frac{Z_n}{2\gamma_n}$$

$$\begin{aligned} \sum_{u=1}^n [f(\theta_{u-1}) - f(\theta_*)] &\leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \sum_{u=1}^n \frac{1}{2\gamma_u} [\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2] + \sum_{u=1}^n \frac{Z_u}{2\gamma_u} \\ &\leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} + \sum_{u=1}^n \frac{Z_u}{2\gamma_u} \leq \frac{2DB}{\sqrt{n}} + \sum_{u=1}^n \frac{Z_u}{2\gamma_u} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}} \end{aligned}$$

- Need to study  $\sum_{u=1}^n \frac{Z_u}{2\gamma_u}$  with  $\mathbb{E}(Z_n | \mathcal{F}_{n-1}) = 0$  and  $|Z_n| \leq 8\gamma_n DB$

# Stochastic subgradient method - high probability - III

- Need to study  $\sum_{u=1}^n \frac{Z_u}{2\gamma_u}$  with  $\mathbb{E}\left(\frac{Z_n}{2\gamma_n} \mid \mathcal{F}_{n-1}\right) = 0$  and  $|Z_n| \leq 4DB$

- Azuma-Hoeffding inequality for bounded martingale increments:

$$\mathbb{P}\left(\sum_{u=1}^n \frac{Z_u}{2\gamma_u} \geq t\sqrt{n} \cdot 4DB\right) \leq \exp\left(-\frac{t^2}{2}\right)$$

- Moments with Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994)

# Beyond stochastic gradient method

- **Adding a proximal step**

- Goal:  $\min_{\theta \in \mathbb{R}^d} f(\theta) + \Omega(\theta) = \mathbb{E}f_n(\theta) + \Omega(\theta)$

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

$$\theta_n = \min_{\theta \in \mathbb{R}^d} \left\| \theta - \theta_{n-1} + \gamma_n f'_n(\theta_n) \right\|_2^2 + C\Omega(\theta)$$

- Xiao (2010); Hu et al. (2009)

- May be accelerated (Ghadimi and Lan, 2013)

- **Related frameworks**

- Regularized dual averaging (Nesterov, 2009; Xiao, 2010)

- Mirror descent (Nemirovski et al., 2009; Lan et al., 2012)

# Mirror descent

- Projected (stochastic) gradient descent adapted to Euclidean geometry

- bound: 
$$\frac{\max_{\theta, \theta' \in \Theta} \|\theta - \theta'\|_2 \cdot \max_{\theta \in \Theta} \|f'(\theta)\|_2}{\sqrt{n}}$$

- What about other norms?

- Example: natural bound on  $\max_{\theta \in \Theta} \|f'(\theta)\|_\infty$  leads to  $\sqrt{d}$  factor
  - Avoidable with **mirror descent**, which leads to factor  $\sqrt{\log d}$
  - Nemirovski et al. (2009); Lan et al. (2012)

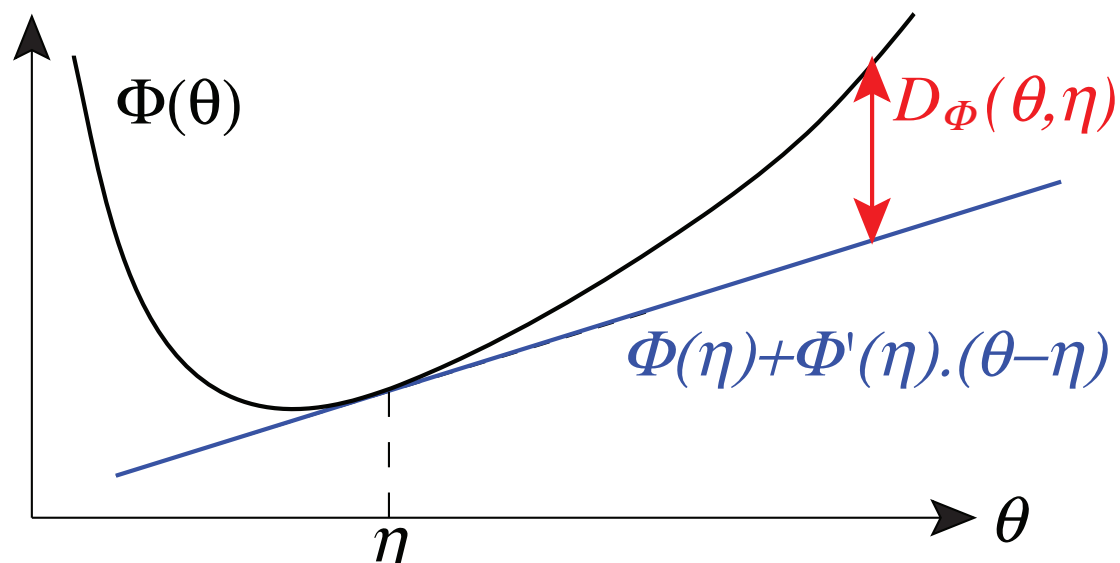
# Mirror descent

- Projected (stochastic) gradient descent adapted to Euclidean geometry
  - bound:  $\frac{\max_{\theta, \theta' \in \Theta} \|\theta - \theta'\|_2 \cdot \max_{\theta \in \Theta} \|f'(\theta)\|_2}{\sqrt{n}}$
- What about other norms?
  - Example: natural bound on  $\max_{\theta \in \Theta} \|f'(\theta)\|_\infty$  leads to  $\sqrt{d}$  factor
  - Avoidable with **mirror descent**, which leads to factor  $\sqrt{\log d}$
  - Nemirovski et al. (2009); Lan et al. (2012)
- From Hilbert to Banach spaces
  - Gradient  $f'(\theta)$  defined through  $f(\theta + d\theta) - f(\theta) = \langle f'(\theta), d\theta \rangle$  for a certain dot-product
  - Generally, the differential is an element of the dual space

## Mirror descent set-up

- Function  $f$  defined on domain  $\mathcal{C}$
- Arbitrary norm  $\|\cdot\|$  with dual norm  $\|s\|_* = \sup_{\|\theta\| \leq 1} \theta^\top s$
- $B$ -Lipschitz-continuous function w.r.t.  $\|\cdot\|$ :  $\|f'(\theta)\|_* \leq B$
- Given a strictly-convex function  $\Phi$ , define the **Bregman divergence**

$$D_\Phi(\theta, \eta) = \Phi(\theta) - \Phi(\eta) - \Phi'(\eta)^\top (\theta - \eta)$$



# Mirror map

- Strongly-convex function  $\Phi : \mathcal{C}_\Phi \rightarrow \mathbb{R}$  such that
  - (a) the gradient  $\Phi'$  takes all possible values in  $\mathbb{R}^d$ , leading to a bijection from  $\mathcal{C}_\Phi$  to  $\mathbb{R}^d$
  - (b) the gradient  $\Phi'$  diverges on the boundary of  $\mathcal{C}_\Phi$
  - (c)  $\mathcal{C}_\Phi$  contains the closure of the domain  $\mathcal{C}$  of the optimization problem
- Bregman projection on  $\mathcal{C}$  uniquely defined on  $\mathcal{C}_\Phi$ :

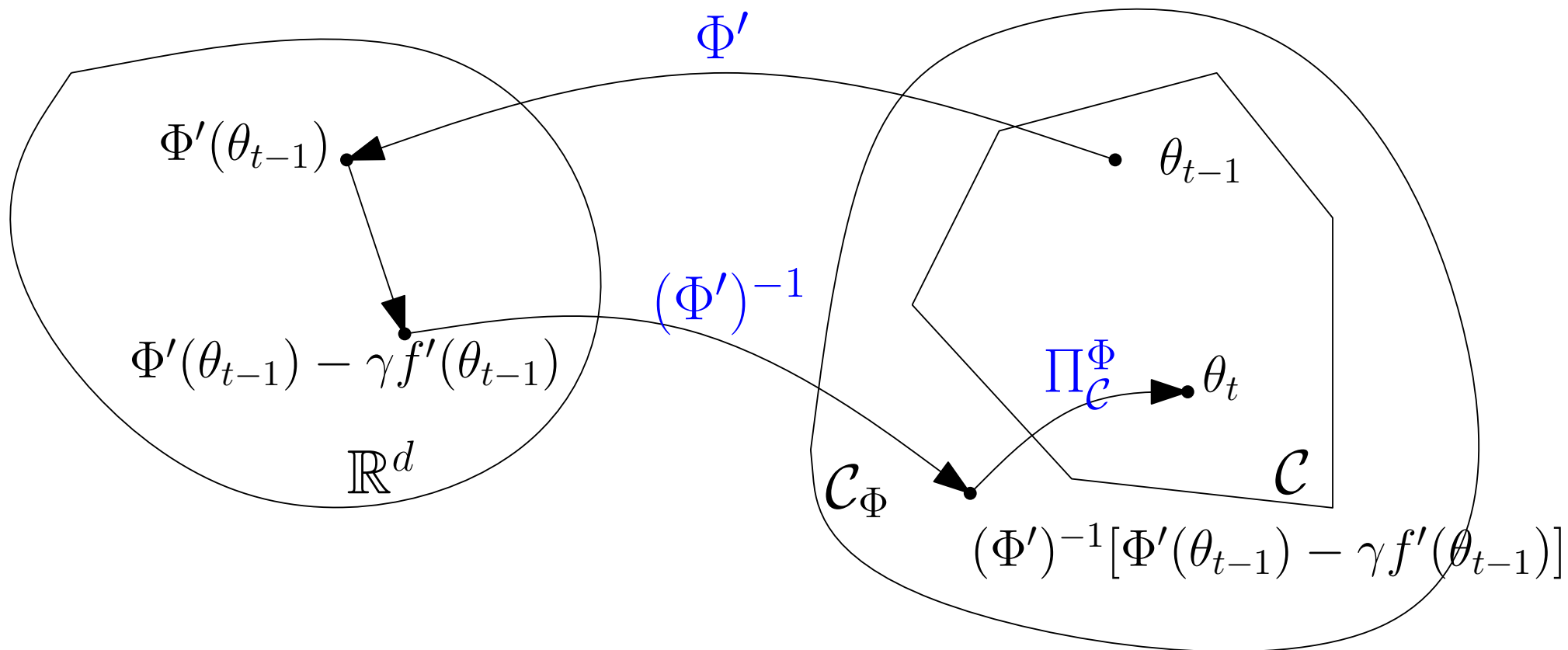
$$\begin{aligned}\Pi_{\mathcal{C}}^{\Phi}(\theta) &= \arg \min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} D_{\Phi}(\eta, \theta) \\ &= \arg \min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi(\theta) - \Phi'(\theta)^{\top}(\eta - \theta) \\ &= \arg \min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi'(\theta)^{\top} \eta\end{aligned}$$

- Example of squared Euclidean norm and entropy

# Mirror descent

- Iteration:

$$\theta_t = \Pi_{\mathcal{C}}^{\Phi}(\Phi'^{-1}[\Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})])$$





# Mirror descent

- **Iteration:**

$$\theta_t = \Pi_{\mathcal{C}}^{\Phi} \left( \Phi'^{-1} \left[ \Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1}) \right] \right)$$

- **Convergence:** assume (a)  $D^2 = \sup_{\theta \in \mathcal{C}} \Phi(\theta) - \inf_{\theta \in \mathcal{C}} \Phi(\theta)$ , (b)  $\Phi$  is  $\alpha$ -strongly convex with respect to  $\|\cdot\|$  and (c)  $f$  is  $B$ -Lipschitz-continuous wr.t.  $\|\cdot\|$ . Then with  $\gamma = \frac{D}{B} \sqrt{\frac{2\alpha}{t}}$ :

$$f\left(\frac{1}{t} \sum_{u=1}^t \theta_u\right) - \inf_{\theta \in \mathcal{C}} f(\theta) \leq DB \sqrt{\frac{2}{\alpha t}}$$

- See detailed proof in Bubeck (2015, p. 299)
- “Same” as subgradient method + allows stochastic gradients

# Mirror descent (proof)

- Define  $\Phi'(\eta_t) = \Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})$ . We have

$$\begin{aligned} f(\theta_{t-1}) - f(\theta) &\leq f'(\theta_{t-1})^\top (\theta_{t-1} - \theta) = \frac{1}{\gamma} (\Phi'(\theta_{t-1}) - \Phi'(\eta_t))^\top (\theta_{t-1} - \theta) \\ &= \frac{1}{\gamma} [D_\Phi(\theta, \theta_{t-1}) + D_\Phi(\theta_{t-1}, \eta_t) - D_\Phi(\theta, \eta_t)] \end{aligned}$$

- By optimality of  $\theta_t$ :  $(\Phi'(\theta_t) - \Phi'(\eta_t))^\top (\theta_t - \theta) \leq 0$  which is equivalent to:  $D_\Phi(\theta, \eta_t) \geq D_\Phi(\theta, \theta_t) + D_\Phi(\theta_t, \eta_t)$ . Thus

$$\begin{aligned} D_\Phi(\theta_{t-1}, \eta_t) - D_\Phi(\theta_t, \eta_t) &= \Phi(\theta_{t-1}) - \Phi(\theta_t) - \Phi'(\eta_t)^\top (\theta_{t-1} - \theta_t) \\ &\leq (\Phi'(\theta_{t-1}) - \Phi'(\eta_t))^\top (\theta_{t-1} - \theta_t) - \frac{\alpha}{2} \|\theta_{t-1} - \theta_t\|^2 \\ &= \gamma f'(\theta_{t-1})^\top (\theta_{t-1} - \theta_t) - \frac{\alpha}{2} \|\theta_{t-1} - \theta_t\|^2 \\ &\leq \gamma B \|\theta_{t-1} - \theta_t\| - \frac{\alpha}{2} \|\theta_{t-1} - \theta_t\|^2 \leq \frac{(\gamma B)^2}{2\alpha} \end{aligned}$$

- Thus  $\sum_{u=1}^t [f(\theta_{t-1}) - f(\theta)] \leq \frac{D_\Phi(\theta, \theta_0)}{\gamma} + \gamma \frac{L^2 t}{2\alpha}$

# Mirror descent examples

- **Euclidean:**  $\Phi = \frac{1}{2} \|\cdot\|_2^2$  with  $\|\cdot\| = \|\cdot\|_2$  and  $\mathcal{C}_\Phi = \mathbb{R}^d$ 
  - Regular gradient descent
- **Simplex:**  $\Phi(\theta) = \sum_{i=1}^d \theta_i \log \theta_i$  with  $\|\cdot\| = \|\cdot\|_1$  and  $\mathcal{C}_\Phi = \{\theta \in \mathbb{R}_+^d, \sum_{i=1}^d \theta_i = 1\}$ 
  - Bregman divergence = Kullback-Leibler divergence
  - Iteration (multiplicative update):  $\theta_t \propto \theta_{t-1} \exp(-\gamma f'(\theta_{t-1}))$
  - Constant:  $D^2 = \log d$ ,  $\alpha = 1$
- **$\ell_p$ -ball:**  $\Phi(\theta) = \frac{1}{2} \|\theta\|_p^2$ , with  $\|\cdot\| = \|\cdot\|_p$ ,  $p \in (1, 2]$ 
  - We have  $\alpha = p - 1$
  - Typically used with  $p = 1 + \frac{1}{\log d}$  to cover the  $\ell_1$ -geometry

# Minimax rates (Agarwal et al., 2012)

- **Model of computation (i.e., algorithms): first-order oracle**
  - Queries a function  $f$  by obtaining  $f(\theta_k)$  and  $f'(\theta_k)$  with zero-mean bounded variance noise, for  $k = 0, \dots, n - 1$  and outputs  $\theta_n$
- **Class of functions**
  - convex  $B$ -Lipschitz-continuous (w.r.t.  $\ell_2$ -norm) on a compact convex set  $\mathcal{C}$  containing an  $\ell_\infty$ -ball
- **Performance measure**
  - for a given algorithm and function  $\varepsilon_n(\text{algo}, f) = f(\theta_n) - \inf_{\theta \in \mathcal{C}} f(\theta)$
  - for a given algorithm: 
$$\sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$$
- **Minimax performance:** 
$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$$

# Minimax rates (Agarwal et al., 2012)

- **Convex functions:** domain  $\mathcal{C}$  that contains an  $\ell_\infty$ -ball of radius  $D$

$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon(\text{algo}, f) \geq \text{cst} \times \min \left\{ BD \sqrt{\frac{d}{n}}, BD \right\}$$

- Consequences for  $\ell_2$ -ball of radius  $D$ :  $BD/\sqrt{n}$
- Upper-bound through stochastic subgradient

- **$\mu$ -strongly-convex functions:**

$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f) \geq \text{cst} \times \min \left\{ \frac{B^2}{\mu n}, \frac{B^2}{\mu d}, BD \sqrt{\frac{d}{n}}, BD \right\}$$

# Minimax rates - sketch of proof

1. **Create a subclass of functions** indexed by some vertices  $\alpha^j$ ,  $j = 1, \dots, M$  of the hypercube  $\{-1, 1\}^d$ , which are sufficiently far in Hamming metric  $\Delta_H$  (denote  $\mathcal{V}$  this set with  $|\mathcal{V}| = M$ )

$$\forall j \neq k, \Delta_H(\alpha^i, \alpha^j) \geq \frac{d}{4},$$

e.g., a “ $\frac{d}{4}$ -packing” (possible with  $M$  exponential in  $d$  - see later)

# Minimax rates - sketch of proof

1. **Create a subclass of functions** indexed by some vertices  $\alpha^j$ ,  $j = 1, \dots, M$  of the hypercube  $\{-1, 1\}^d$ , which are sufficiently far in Hamming metric  $\Delta_H$  (denote  $\mathcal{V}$  this set with  $|\mathcal{V}| = M$ )

$$\forall j \neq k, \Delta_H(\alpha^i, \alpha^j) \geq \frac{d}{4},$$

e.g., a “ $\frac{d}{4}$ -packing” (possible with  $M$  exponential in  $d$  - see later)

2. **Design functions** so that

- approximate optimization of the function is equivalent to function identification among the class above
- stochastic oracle corresponds to a sequence of coin tosses with biases index by  $\alpha^j$ ,  $j = 1, \dots, M$

# Minimax rates - sketch of proof

1. **Create a subclass of functions** indexed by some vertices  $\alpha^j$ ,  $j = 1, \dots, M$  of the hypercube  $\{-1, 1\}^d$ , which are sufficiently far in Hamming metric  $\Delta_H$  (denote  $\mathcal{V}$  this set with  $|\mathcal{V}| = M$ )

$$\forall j \neq k, \Delta_H(\alpha^i, \alpha^j) \geq \frac{d}{4},$$

e.g., a “ $\frac{d}{4}$ -packing” (possible with  $M$  exponential in  $d$  - see later)

2. **Design functions** so that

- approximate optimization of the function is equivalent to function identification among the class above
- stochastic oracle corresponds to a sequence of coin tosses with biases index by  $\alpha^j$ ,  $j = 1, \dots, M$

3. Any such identification procedure (i.e., **a test**) has a lower bound on the probability of error



# Packing number for the hyper-cube

## Proof

- **Varshamov-Gilbert's lemma** (Massart, 2003, p. 105): the maximal number of points in the hypercube that are at least  $d/4$ -apart in Hamming loss is greater than  $\exp(d/8)$ .

1. Maximality of family  $\mathcal{V} \Rightarrow \bigcup_{\alpha \in \mathcal{V}} \mathcal{B}_H(\alpha, d/4) = \{-1, 1\}^d$

2. Cardinality:  $\sum_{\alpha \in \mathcal{V}} |\mathcal{B}_H(\alpha, d/4)| \geq 2^d$

3. Link with deviation of  $Z$  distributed as Binomial( $d, 1/2$ )

$$2^{-d} |\mathcal{B}_H(\alpha, d/4)| = \mathbb{P}(Z \leq d/4) = \mathbb{P}(Z \geq 3d/4)$$

4. Hoeffding inequality:  $\mathbb{P}(Z - \frac{d}{2} \geq \frac{d}{4}) \leq \exp(-\frac{2(d/4)^2}{d}) = \exp(-\frac{d}{8})$

# Designing a class of functions

- Given  $\alpha \in \{-1, 1\}^d$ , and a precision parameter  $\delta > 0$ :

$$g_\alpha(x) = \frac{c}{d} \sum_{i=1}^d \left\{ \left( \frac{1}{2} + \alpha_i \delta \right) f_i^+(x) + \left( \frac{1}{2} - \alpha_i \delta \right) f_i^-(x) \right\}$$

- **Properties**

- Functions  $f_i$ 's and constant  $c$  to ensure proper regularity and/or strong convexity

- **Oracle**

- (a) Pick an index  $i \in \{1, \dots, d\}$  at random
- (b) Draw  $b_i \in \{0, 1\}$  from a Bernoulli with parameter  $\frac{1}{2} + \alpha_i \delta$
- (c) Consider  $\hat{g}_\alpha(x) = c[b_i f_i^+ + (1 - b_i) f_i^-]$  and its value / gradient

# Optimizing is function identification

- **Goal:** if  $g_\alpha$  is optimized up to error  $\varepsilon$ , then this identifies  $\alpha \in \mathcal{V}$

- “Metric” between functions:

$$\rho(f, g) = \inf_{\theta \in \mathcal{C}} f(\theta) + g(\theta) - \inf_{\theta \in \mathcal{C}} f(\theta) - \inf_{\theta \in \mathcal{C}} g(\theta)$$

–  $\rho(f, g) \geq 0$  with equality iff  $f$  and  $g$  have the same minimizers

- **Lemma:** let  $\psi(\delta) = \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta)$ . For any  $\tilde{\theta} \in \mathcal{C}$ , there is at most one function  $g_\alpha$  such that  $g_\alpha(\tilde{\theta}) - \inf_{\theta \in \mathcal{C}} g_\alpha(\theta) \leq \frac{\psi(\delta)}{3}$

# Optimizing is function identification

- **Goal:** if  $g_\alpha$  is optimized up to error  $\varepsilon$ , then this identifies  $\alpha \in \mathcal{V}$

- **“Metric” between functions:**

$$\rho(f, g) = \inf_{\theta \in \mathcal{C}} f(\theta) + g(\theta) - \inf_{\theta \in \mathcal{C}} f(\theta) - \inf_{\theta \in \mathcal{C}} g(\theta)$$

- $\rho(f, g) \geq 0$  with equality iff  $f$  and  $g$  have the same minimizers

- **Lemma:** let  $\psi(\delta) = \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta)$ . For any  $\tilde{\theta} \in \mathcal{C}$ , there is at most one function  $g_\alpha$  such that  $g_\alpha(\tilde{\theta}) - \inf_{\theta \in \mathcal{C}} g_\alpha(\theta) \leq \frac{\psi(\delta)}{3}$

- (a) optimizing an unknown function from the class up to precision  $\frac{\psi(\delta)}{3}$  leads to identification of  $\alpha \in \mathcal{V}$

- (b) If the expected minimax error rate is greater than  $\frac{\psi(\delta)}{9}$ , there exists a function from the set of random gradient and function values such the probability of error is less than  $1/3$

## Lower bounds on coin tossing (Agarwal et al., 2012, Lemma 3)

- **Lemma:** For  $\delta < 1/4$ , given  $\alpha^*$  uniformly at random in  $\mathcal{V}$ , if  $n$  outcomes of a random single coin (out of the  $d$ ) are revealed, then any test will have a probability of error greater than

$$1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2} \log(2/\sqrt{e})}$$

- Proof based on Fano's inequality: If  $g$  is a function of  $Y$ , and  $X$  takes  $m$  values, then

$$\mathbb{P}(g(X) \neq Y) \geq \frac{H(X|Y) - 1}{\log m} = \frac{H(X)}{\log m} - \frac{I(X, Y) + 1}{\log m}$$

# Construction of $f_i$ for convex functions

- $f_i^+(\theta) = |\theta(i) + \frac{1}{2}|$  and  $f_i^-(\theta) = |\theta(i) - \frac{1}{2}|$ 
  - 1-Lipschitz-continuous with respect to the  $\ell_2$ -norm. With  $c = B/2$ , then  $g_\alpha$  is  $B$ -Lipschitz.
  - Calling the oracle reveals a coin
- Lower bound on the discrepancy function
  - each  $g_\alpha$  is minimized at  $\theta_\alpha = -\alpha/2$
  - Fact:  $\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d} \Delta_H(\alpha, \beta) \geq \frac{c\delta}{2} = \psi(\delta)$
- Set error/precision  $\varepsilon = \frac{c\delta}{18}$  so that  $\varepsilon < \psi(\delta)/9$
- Consequence:  $\frac{1}{3} \geq 1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2} \log(2/\sqrt{e})}$ , that is,  $n \geq \text{cst} \times \frac{L^2 d^2}{\varepsilon^2}$

## Construction of $f_i$ for strongly-convex functions

- $f_i^\pm(\theta) = \frac{1}{2}\kappa|\theta(i) \pm \frac{1}{2}| + \frac{1-\kappa}{4}(\theta(i) \pm \frac{1}{2})^2$ 
  - Strongly convex and Lipschitz-continuous
- Same proof technique (more technical details)
- See more details by Agarwal et al. (2012); Raginsky and Rakhlin (2011)

# Summary of rates of convergence

- Problem parameters
  - $D$  diameter of the domain
  - $B$  Lipschitz-constant
  - $L$  smoothness constant
  - $\mu$  strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$ stochastic: $BD/\sqrt{n}$	deterministic: $B^2/(t\mu)$ stochastic: $B^2/(n\mu)$
smooth	deterministic: $LD^2/t^2$	deterministic: $\exp(-t\sqrt{\mu/L})$
quadratic	deterministic: $LD^2/t^2$	deterministic: $\exp(-t\sqrt{\mu/L})$



# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# “Classical” stochastic approximation

- **General problem of finding zeros of  $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$** 
  - From random observations of values of  $h$  at certain points
  - Main example: minimization of  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , with  $h = f'$
- **Classical algorithm (Robbins and Monro, 1951b)**

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

# “Classical” stochastic approximation

- **General problem of finding zeros of  $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$** 
  - From random observations of values of  $h$  at certain points
  - Main example: minimization of  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , with  $h = f'$

- **Classical algorithm (Robbins and Monro, 1951b)**

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

- **Goals** (see, e.g., Duflo, 1996)
  - Beyond reducing noise by averaging observations
  - General sufficient conditions for convergence
  - Convergence in quadratic mean vs. convergence almost surely
  - Rates of convergences and choice of step-sizes
  - Asymptotics - no convexity

# “Classical” stochastic approximation

- Intuition from recursive mean estimation

- Starting from  $\theta_0 = 0$ , getting data  $x_n \in \mathbb{R}^d$

$$\theta_n = \theta_{n-1} - \gamma_n(\theta_{n-1} - x_n)$$

- If  $\gamma_n = 1/n$ , then  $\theta_n = \frac{1}{n} \sum_{k=1}^n x_k$
- If  $\gamma_n = 2/(n+1)$  then  $\theta_n = \frac{2}{n(n+1)} \sum_{k=1}^n kx_k$

# “Classical” stochastic approximation

- Intuition from recursive mean estimation

- Starting from  $\theta_0 = 0$ , getting data  $x_n \in \mathbb{R}^d$

$$\theta_n = \theta_{n-1} - \gamma_n(\theta_{n-1} - x_n)$$

- If  $\gamma_n = 1/n$ , then  $\theta_n = \frac{1}{n} \sum_{k=1}^n x_k$

- If  $\gamma_n = 2/(n+1)$  then  $\theta_n = \frac{2}{n(n+1)} \sum_{k=1}^n kx_k$

- In general:  $\mathbb{E}x_n = x$  and thus  $\theta_n - x = (1 - \gamma_n)(\theta_{n-1} - x) + \gamma_n(x_n - x)$

$$\theta_n - x = \prod_{k=1}^n (1 - \gamma_k)(\theta_0 - x) + \sum_{i=1}^n \prod_{k=i+1}^n (1 - \gamma_k) \gamma_i (x_i - x)$$

# “Classical” stochastic approximation

- Expanding the recursion with i.i.d.  $x_n$ 's and  $\sigma^2 = \mathbb{E}\|x_n - x\|^2$ :

$$\theta_n - x = \prod_{k=1}^n (1 - \gamma_k)(\theta_0 - x) + \sum_{i=1}^n \gamma_i \prod_{k=i+1}^n (1 - \gamma_k)(x_i - x)$$

$$\mathbb{E}\|\theta_n - x\|^2 = \prod_{k=1}^n (1 - \gamma_k)^2 \|\theta_0 - x\|^2 + \sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - \gamma_k)^2 \sigma^2$$

# “Classical” stochastic approximation

- Expanding the recursion with i.i.d.  $x_n$ 's and  $\sigma^2 = \mathbb{E}\|x_n - x\|^2$ :

$$\theta_n - x = \prod_{k=1}^n (1 - \gamma_k)(\theta_0 - x) + \sum_{i=1}^n \gamma_i \prod_{k=i+1}^n (1 - \gamma_k)(x_i - x)$$

$$\mathbb{E}\|\theta_n - x\|^2 = \prod_{k=1}^n (1 - \gamma_k)^2 \|\theta_0 - x\|^2 + \sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - \gamma_k)^2 \sigma^2$$

- Requires study of  $\prod_{k=1}^n (1 - \gamma_k)$  and  $\sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - \gamma_k)^2$ 
  - If  $\gamma_n = o(1)$ ,  $\log \prod_{k=1}^n (1 - \gamma_k) \sim -\sum_{k=1}^n \gamma_k$  should go to  $-\infty$   
**Forgetting initial conditions (even arbitrarily far)**
  - $\sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - \gamma_k)^2 \sim \sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - 2\gamma_k)$   
**Robustness to noise**



# Forgetting of initial conditions

$$\log \prod_{k=1}^n (1 - \gamma_k) \sim - \sum_{k=1}^n \gamma_k$$

- Examples:  $\boxed{\gamma_n = C/n^\alpha}$

- $\alpha = 1$ ,  $\sum_{i=1}^n \frac{1}{i} = \log(n) + \text{cst} + O(1/n)$
- $\alpha > 1$ ,  $\sum_{i=1}^n \frac{1}{i^\alpha} = \text{cst} + O(1/n^{\alpha-1})$
- $\alpha \in (0, 1)$ ,  $\sum_{i=1}^n \frac{1}{i^\alpha} = \text{cst} \times n^{1-\alpha} + O(1)$
- Proof using relationship with integrals

- Consequences

- if  $\alpha > 1$ , no convergence
- If  $\alpha \in (0, 1)$ , exponential convergence
- if  $\alpha = 1$ , convergence of squared norm in  $1/n^{2C}$

## Decomposition of the noise term

- Assume  $(\gamma_n)$  is decreasing and less than  $1/\mu$ ; then for any  $m \in \{1, \dots, n\}$ , we may split the following sum as follows:

$$\begin{aligned}
 \sum_{k=1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 &= \sum_{k=1}^m \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 + \sum_{k=m+1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 \\
 &\leq \prod_{i=m+1}^n (1 - \mu\gamma_i) \sum_{k=1}^m \gamma_k^2 + \gamma_m \sum_{k=m+1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k \\
 &\leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^m \gamma_k^2 + \frac{\gamma_m}{\mu} \sum_{k=m+1}^n \left[ \prod_{i=k+1}^n (1 - \mu\gamma_i) - \prod_{i=k}^n (1 - \mu\gamma_i) \right] \\
 &\leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^m \gamma_k^2 + \frac{\gamma_m}{\mu} \left[ 1 - \prod_{i=m+1}^n (1 - \mu\gamma_i) \right] \\
 &\leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^n \gamma_k^2 + \frac{\gamma_m}{\mu}
 \end{aligned}$$

## Decomposition of the noise term

$$\sum_{k=1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 \leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^n \gamma_k^2 + \frac{\gamma_m}{\mu}$$

- Require  $\gamma_n$  to tend to zero (vanishing decaying step-size)
  - May not need  $\sum_n \gamma_n^2 < \infty$  for convergence in quadratic mean
- Examples:  $\boxed{\gamma_n = C/n^\alpha}$  and mean estimation ( $\mu = 1$ )
  - No need to consider  $\alpha > 1$
  - $\alpha \in (0, 1)$ ,
  - $\alpha = 1$ , convergence of noise term in  $O(1/n)$  but forgetting of initial condition in  $O(1/n^{2C})$
  - Consequences: **need  $\alpha \in (0, 1]$**  and  $C \geq 1/2$  for  $\alpha = 1$

# Robbins-Monro algorithm

- **General problem of finding zeros of  $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$** 
  - From random observations of values of  $h$  at certain points
  - Main example: minimization of  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , with  $h = f'$

- **Classical algorithm (Robbins and Monro, 1951b)**

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

- **Goals** (see, e.g., Duflo, 1996)
  - General sufficient conditions for convergence
  - Convergence in quadratic mean vs. convergence almost surely
  - Rates of convergences and choice of step-sizes
  - Asymptotics - no convexity

# Different types of convergences

- **Goal:** show that  $\theta_n \rightarrow \theta_*$  or  $d(\theta_n, \Theta_*) \rightarrow 0$  or  $f(\theta_n) \rightarrow f(\theta_*)$ 
  - Random quantity  $\delta_n \in \mathbb{R}$  tending to zero
- **Convergence almost-surely:**  $\mathbb{P}(\delta_n \rightarrow 0) = 1$
- **Convergence in probability:**  $\forall \varepsilon > 0, \mathbb{P}(|\delta_n| \geq \varepsilon) \rightarrow 0$
- **Convergence in mean**  $r \geq 1$ :  $\mathbb{E}|\delta_n|^r \rightarrow 0$

# Different types of convergences

- **Goal:** show that  $\theta_n \rightarrow \theta_*$  or  $d(\theta_n, \Theta_*) \rightarrow 0$  or  $f(\theta_n) \rightarrow f(\theta_*)$ 
  - Random quantity  $\delta_n \in \mathbb{R}$  tending to zero
- **Convergence almost-surely:**  $\mathbb{P}(\delta_n \rightarrow 0) = 1$
- **Convergence in probability:**  $\forall \varepsilon > 0, \mathbb{P}(|\delta_n| \geq \varepsilon) \rightarrow 0$
- **Convergence in mean**  $r \geq 1$ :  $\mathbb{E}|\delta_n|^r \rightarrow 0$
- **Relationship between convergences**
  - Almost surely  $\Rightarrow$  in probability
  - In mean  $\Rightarrow$  in probability (Markov's inequality)
  - In probability (sufficiently fast)  $\Rightarrow$  almost surely (Borel-Cantelli)
  - Almost surely + domination  $\Rightarrow$  in mean

# Robbins-Monro algorithm

## Need for Lyapunov functions (even with no noise)

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

- The Robbins-Monro algorithm cannot converge all the time...
- **Lyapunov function**  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  with following properties
  - Non-negative values:  $V \geq 0$
  - Continuously-differentiable with  $L$ -Lipschitz-continuous gradients
  - Control of  $h$ :  $\forall \theta, \|h(\theta)\|^2 \leq C(1 + V(\theta))$
  - Gradient condition:  $\forall \theta, \boxed{h(\theta)^\top V'(\theta) \geq \alpha \|V'(\theta)\|^2}$

# Robbins-Monro algorithm

## Need for Lyapunov functions (even with no noise)

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

- The Robbins-Monro algorithm cannot converge all the time...
- **Lyapunov function**  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  with following properties
  - Non-negative values:  $V \geq 0$
  - Continuously-differentiable with  $L$ -Lipschitz-continuous gradients
  - Control of  $h$ :  $\forall \theta, \|h(\theta)\|^2 \leq C(1 + V(\theta))$
  - Gradient condition:  $\forall \theta, \boxed{h(\theta)^\top V'(\theta) \geq \alpha \|V'(\theta)\|^2}$
- If  $h = f'$ , then  $V(\theta) = f(\theta) - \inf f$  is the default (but not only) choice for Lyapunov function: **applies also to non-convex functions**
  - Will require often some additional condition  $\|V'(\theta)\|^2 \geq 2\mu V(\theta)$



# Robbins-Monro algorithm

## Martingale noise

$$\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$$

- **Assumptions about the noise**  $\varepsilon_n$

- Typical assumption:  $\varepsilon_n$  i.i.d.  $\Rightarrow$  **not needed**
- “information up to time  $n$ ”: sequence of increasing  $\sigma$ -fields  $\mathcal{F}_n$
- Example from machine learning:  $\mathcal{F}_n = \sigma(x_1, y_1, \dots, x_n, y_n)$
- Assume  $\mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0$  and  $\mathbb{E}[\|\varepsilon_n\|^2 | \mathcal{F}_{n-1}] \leq \sigma^2$  almost surely

- **Key property:**  $\theta_n$  is  $\mathcal{F}_n$ -measurable

# Robbins-Monro algorithm

## Convergence of the Lyapunov function

- Using regularity (and other properties) of  $V$ :

$$\begin{aligned} V(\theta_n) &\leq V(\theta_{n-1}) + V'(\theta_{n-1})^\top (\theta_n - \theta_{n-1}) + \frac{L}{2} \|\theta_n - \theta_{n-1}\|^2 \\ &= V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top (h(\theta_{n-1}) + \varepsilon_n) + \frac{L\gamma_n^2}{2} \|h(\theta_{n-1}) + \varepsilon_n\|^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[V(\theta_n) | \mathcal{F}_{n-1}] &\leq V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top h(\theta_{n-1}) + \frac{L\gamma_n^2}{2} \|h(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2} \sigma^2 \\ &\leq V(\theta_{n-1}) - \alpha\gamma_n \|V'(\theta_{n-1})\|^2 + \frac{LC\gamma_n^2}{2} [1 + V(\theta_{n-1})] + \frac{L\gamma_n^2}{2} \sigma^2 \\ &\leq V(\theta_{n-1}) \left[1 + \frac{LC\gamma_n^2}{2}\right] - \alpha\gamma_n \|V'(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2} (C + \sigma^2) \end{aligned}$$

# Robbins-Monro algorithm

## Convergence of the expected Lyapunov function with “curvature”

- If  $\|V'(\theta)\|^2 \geq 2\mu V(\theta)$  and  $\gamma_n \leq \frac{2\alpha\mu}{LC}$ :

$$\mathbb{E}[V(\theta_n)|\mathcal{F}_{n-1}] \leq V(\theta_{n-1})[1 - \alpha\mu\gamma_n] + M\gamma_n^2$$

$$\mathbb{E}V(\theta_n) \leq \mathbb{E}V(\theta_{n-1})[1 - \alpha\mu\gamma_n] + M\gamma_n^2$$

- Need to study non-negative sequence  $\delta_n \leq \delta_{n-1}[1 - \alpha\mu\gamma_n] + M\gamma_n^2$  with  $\delta_n = \mathbb{E}V(\theta_n)$
- Sufficient conditions for convergence of the expected Lyapunov function (with curvature)
  - $\sum_n \gamma_n = +\infty$  and  $\gamma_n \rightarrow 0$
  - Special case of  $\gamma_n = C/n^\alpha$

# Robbins-Monro algorithm

## Convergence of the expected Lyapunov function

with “curvature” -  $\gamma_n = C/n^\alpha$

- Need to study non-negative sequence  $\delta_n \leq \delta_{n-1} [1 - \alpha\mu\gamma_n] + M\gamma_n^2$  with  $\delta_n = \mathbb{E}V(\theta_n)$  (NB: forgetting constraint on  $\gamma_n$  - see next class)

$$\delta_n \leq \prod_{k=1}^n (1 - \alpha\mu\gamma_k) \delta_0 + M \sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - \alpha\mu\gamma_k)$$

- If  $\alpha > 1$ : no forgetting of initial conditions
- If  $\alpha \in (0, 1)$ :  $\delta_0 \exp(-\text{cst } \alpha\mu C \times n^{\alpha-1}) + \gamma_n M$
- If  $\alpha = 1$  and  $\gamma_n = C/n$ :  $\delta_0 n^{-\mu C} + \gamma_n M$

# Robbins-Monro algorithm

## Convergence of the expected Lyapunov function

with “curvature” -  $\gamma_n = C/n^\alpha$

- Summary of the rates with dependence on noise

# Robbins-Monro algorithm

## Almost-sure convergence

- Using regularity of  $V$ :

$$\begin{aligned} V(\theta_n) &\leq V(\theta_{n-1}) + V'(\theta_{n-1})^\top (\theta_n - \theta_{n-1}) + \frac{L}{2} \|\theta_n - \theta_{n-1}\|^2 \\ &= V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top (h(\theta_{n-1}) + \varepsilon_n) + \frac{L\gamma_n^2}{2} \|h(\theta_{n-1}) + \varepsilon_n\|^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[V(\theta_n) | \mathcal{F}_{n-1}] &\leq V(\theta_{n-1}) - \gamma_n V'(\theta_{n-1})^\top h(\theta_{n-1}) + \frac{L\gamma_n^2}{2} \|h(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2} \sigma^2 \\ &\leq V(\theta_{n-1}) - \alpha\gamma_n \|V'(\theta_{n-1})\|^2 + \frac{LC\gamma_n^2}{2} [1 + V(\theta_{n-1})] + \frac{L\gamma_n^2}{2} \sigma^2 \\ &= V(\theta_{n-1}) \left[1 + \frac{LC\gamma_n^2}{2}\right] - \alpha\gamma_n \|V'(\theta_{n-1})\|^2 + \frac{L\gamma_n^2}{2} (C + \sigma^2) \end{aligned}$$

# Robbins and Siegmund (1985)

- **Assumptions**

- Measurability: Let  $V_n, \beta_n, \chi_n, \eta_n$  four  $\mathcal{F}_n$ -adapted real sequences
- Non-negativity:  $V_n, \beta_n, \chi_n, \eta_n$  non-negative
- Summability:  $\sum_n \beta_n < \infty$  and  $\sum_n \chi_n < \infty$
- Inequality:  $\mathbb{E}[V_n | \mathcal{F}_{n-1}] \leq V_{n-1}(1 + \beta_{n-1}) + \chi_{n-1} - \eta_{n-1}$

- **Theorem:**  $(V_n)$  converges almost surely to a random variable  $V_\infty$  and  $\sum_n \eta_n$  is finite almost surely

- *Proof*

- Consequence for stochastic approximation (if  $\|V'(\theta)\|^2 \geq 2\mu V(\theta)$ ):  $V(\theta_n)$  and  $\|V'(\theta_n)\|^2$  converges almost surely to zero

# Robbins and Siegmund (1985) - Proof sketch

- Inequality:  $\mathbb{E}[V_n | \mathcal{F}_{n-1}] \leq V_{n-1}(1 + \beta_{n-1}) + \chi_{n-1} - \eta_{n-1}$
- Define  $\alpha_n = \prod_{k=1}^n (1 + \beta_k)$  a converging sequence,  $V'_n = \alpha_{n-1} V_n$ ,  $\chi'_n = \alpha_{n-1} \chi_n$  and  $\eta'_n = \alpha_{n-1} \eta_n$  so that:

$$\mathbb{E}[V'_n | \mathcal{F}_{n-1}] \leq V_{n-1} + \chi'_{n-1} - \eta'_{n-1}$$

- Define the super-martingale  $Y_n = V'_n - \sum_{k=1}^{n-1} (\chi'_k - \eta'_k)$  so that

$$\mathbb{E}[Y_n | \mathcal{F}_{n-1}] \leq Y_{n-1}$$

- Probabilistic proof using Doob convergence theorem (Duflo, 1996)



# Robbins-Monro analysis - non random errors

- **Random unbiased errors:** no need for vanishing magnitudes
- **Non-random errors:** need for vanishing magnitudes
  - See Duflo (1996, Theorem 2.III.4)
  - See also Schmidt et al. (2011)

# Robbins-Monro analysis - asymptotic normality (Fabian, 1968)

- Traditional step-size  $\gamma = C/n$  (and proof sketch for differential  $A$  of  $h$  at unique  $\theta_*$  symmetric)

$$\begin{aligned}\theta_n &= \theta_{n-1} - \gamma_n h(\theta_{n-1}) - \gamma_n \varepsilon_n \\ &\approx \theta_{n-1} - \gamma_n [h'(\theta_*)(\theta_{n-1} - \theta_*)] - \gamma_n \varepsilon_n + \gamma_n O(\|\theta_n - \theta_*\|^2) \\ &\approx \theta_{n-1} - \gamma_n A(\theta_{n-1} - \theta_*) - \gamma_n \varepsilon_n\end{aligned}$$

$$\theta_n - \theta_* \approx (I - \gamma_n A) \cdots (I - \gamma_1 A)(\theta_0 - \theta_*) - \sum_{k=1}^n (I - \gamma_n A) \cdots (I - \gamma_{k+1} A) \gamma_k \varepsilon_k$$

$$\theta_n - \theta_* \approx \exp[-(\gamma_n + \cdots + \gamma_1)A](\theta_0 - \theta_*) - \sum_{k=1}^n \exp[-(\gamma_n + \cdots + \gamma_{k+1})A] \gamma_k \varepsilon_k$$

$$\approx \exp[-CA \log n](\theta_0 - \theta_*) - \sum_{k=1}^n \exp[-C(\log n - \log k)A] \frac{C}{k} \varepsilon_k$$

- Asymptotic normality by averaging random variables

# Robbins-Monro analysis - asymptotic normality (Fabian, 1968)

- Assuming  $A$ ,  $(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top$  and  $\mathbb{E}(\varepsilon_k \varepsilon_k^\top) = \Sigma$  commute

$$\theta_n - \theta_* \approx \exp[-CA \log n](\theta_0 - \theta_*) - \sum_{k=1}^n \exp[-C(\log n - \log k)A] \frac{C}{k} \varepsilon_k$$

$$\begin{aligned} \mathbb{E}(\theta_n - \theta_*)(\theta_n - \theta_*)^\top &\approx \exp[-2CA \log n](\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top \\ &\quad + \sum_{k=1}^n \exp[-2C(\log n - \log k)A] \frac{C^2}{k^2} \mathbb{E}(\varepsilon_k \varepsilon_k^\top) \\ &\approx n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + n^{-2CA} \sum_{k=1}^n C^2 k^{2CA-2} \Sigma \\ &\approx n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + n^{-2CA} C^2 \frac{n^{2CA-1}}{2CA-1} \Sigma \end{aligned}$$

# Robbins-Monro analysis - asymptotic normality (Fabian, 1968)

$$\mathbb{E}(\theta_n - \theta_*)(\theta_n - \theta_*)^\top \approx n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^\top + \frac{1}{n}C^2 \frac{1}{2CA - 1} \Sigma$$

- Step-size  $\gamma = C/n$  (note that this only a sketch of proof)
  - Need  $2C\lambda_{\min}(A) \geq 1$  for convergence, which implies that the first term depending on initial condition  $\theta_* - \theta_0$  is negligible
  - $C$  too small  $\Rightarrow$  no convergence -  $C$  too large  $\Rightarrow$  large variance
- Dependence on the conditioning of the problem
  - If  $\lambda_{\min}(A)$  is small, then  $C$  is large
  - “Choosing”  $A$  proportional to identity for optimal behavior (by premultiplying  $A$  by a conditioning matrix that make  $A$  close to a constant times identity)

# Polyak-Ruppert averaging

- **Problems with Robbins-Monro algorithm**

- Choice of step-sizes in Robbins-Monro algorithm
- Dependence on the unknown conditioning of the problem

- **Simple but impactful idea** (Polyak and Juditsky, 1992; Ruppert, 1988)

- Consider the averaged iterate

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

- NB: “Offline” averaging
- Can be computed recursively as  $\bar{\theta}_n = (1 - 1/n)\bar{\theta}_{n-1} + \frac{1}{n}\theta_n$
- In practice, may start the averaging “after a while”

- **Analysis**

- Unique optimum  $\theta_*$ . See details by Polyak and Juditsky (1992)

# Cesaro means

- Assume  $\theta_n \rightarrow \theta_*$ , with convergence rate  $\|\theta_n - \theta_*\| \leq \alpha_n$
- Cesaro's theorem:  $\bar{\theta}_n = \frac{1}{n} \sum_{k=1}^n \theta_k$  converges to  $\theta_*$
- What about convergence rate  $\|\bar{\theta}_n - \theta_*\|$ ?

# Cesaro means

- Assume  $\theta_n \rightarrow \theta_*$ , with convergence rate  $\|\theta_n - \theta_*\| \leq \alpha_n$
- Cesaro's theorem:  $\bar{\theta}_n = \frac{1}{n} \sum_{k=1}^n \theta_k$  converges to  $\theta_*$
- What about convergence rate  $\|\bar{\theta}_n - \theta_*\|$ ?

$$\|\bar{\theta}_n - \theta_*\| \leq \frac{1}{n} \sum_{k=1}^n \|\theta_k - \theta_*\| \leq \frac{1}{n} \sum_{k=1}^n \alpha_k$$

- Will depend on rate  $\alpha_n$
- If  $\sum_n \alpha_n < \infty$ , the rate becomes  $1/n$  independently of  $\alpha_n$

# Polyak-Ruppert averaging - Proof sketch - I

- Recursion:  $\theta_n = \theta_{n-1} - \gamma_n(h(\theta_{n-1}) + \varepsilon_n)$  with  $\gamma_n = C/n^\alpha$ 
  - From before, we know that  $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$

$$h(\theta_{n-1}) = \frac{1}{\gamma_n} [\theta_{n-1} - \theta_n] - \varepsilon_n$$

$$A(\theta_{n-1} - \theta_*) + O(\|\theta_{n-1} - \theta_*\|^2) = \frac{1}{\gamma_n} [\theta_{n-1} - \theta_n] - \varepsilon_n \text{ with } A = h'(\theta_*)$$

$$A(\theta_{n-1} - \theta_*) = \frac{1}{\gamma_n} [\theta_{n-1} - \theta_n] - \varepsilon_n + O(n^{-\alpha})$$

$$\frac{1}{n} \sum_{k=1}^n A(\theta_{k-1} - \theta_*) = \frac{1}{n} \sum_{k=1}^n \frac{1}{\gamma_k} [\theta_{k-1} - \theta_k] - \frac{1}{n} \sum_{k=1}^n \varepsilon_k + O(n^{-\alpha})$$

$$\frac{1}{n} \sum_{k=1}^n A(\theta_{k-1} - \theta_*) = \frac{1}{n} \sum_{k=1}^n \frac{1}{\gamma_k} [\theta_{k-1} - \theta_k] + \text{Normal}(0, \Sigma/n) + O(n^{-\alpha})$$



# Polyak-Ruppert averaging - Proof sketch - II

- **Goal:** Bounding  $\frac{1}{n} \sum_{k=1}^n \frac{1}{\gamma_k} [\theta_{k-1} - \theta_k]$  given  $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$
- Abel's summation formula We have, summing by parts,

$$\frac{1}{n} \sum_{k=1}^n \frac{1}{\gamma_k} (\theta_{k-1} - \theta_k) = \frac{1}{n} \sum_{k=1}^{n-1} (\theta_k - \theta_*) (\gamma_{k+1}^{-1} - \gamma_k^{-1}) - \frac{1}{n} (\theta_n - \theta_*) \gamma_n^{-1} + \frac{1}{n} (\theta_0 - \theta_*) \gamma_1^{-1}$$

leading to

$$\left\| \frac{1}{n} \sum_{k=1}^n \frac{1}{\gamma_k} (\theta_{k-1} - \theta_k) \right\| \leq \frac{1}{n} \sum_{k=1}^{n-1} \|\theta_k - \theta_*\| \cdot |\gamma_{k+1}^{-1} - \gamma_k^{-1}| + \frac{1}{n} \|\theta_n - \theta_*\| \gamma_n^{-1} + \frac{1}{n} \|\theta_0 - \theta_*\| \gamma_1^{-1}$$

which is negligible

# Polyak-Ruppert averaging - Proof sketch - III

- Recursion:  $\theta_n = \theta_{n-1} - \gamma_n(h(\theta_{n-1}) + \varepsilon_n)$  with  $\gamma_n = C/n^\alpha$ 
  - From before, we know that  $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$

$$\frac{1}{n} \sum_{k=1}^n A(\theta_{k-1} - \theta_*) = \text{Normal}(0, \Sigma/n) + O(n^{-\alpha}) + O(n^{2\alpha-1})$$

- **Consequence:**  $\bar{\theta}_n - \theta_*$  is asymptotically normal with mean zero and covariance  $\frac{1}{n} A^{-1} \Sigma A^{-1}$ 
  - Achieves the Cramer-Rao lower bound (see next lecture)
  - Independent of step-size (see next lecture)
  - Where are the initial conditions? (see next lecture)

# Beyond the classical analysis

- **Lack of strong-convexity**
  - Step-size  $\gamma_n = 1/n$  not robust to ill-conditioning
- **Robustness of step-sizes**
- **Explicit forgetting of initial conditions**

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# Convex stochastic approximation

## Existing work

- Known **global** minimax rates of convergence for **non-smooth** problems (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
  - **Strongly convex:**  $O((\mu n)^{-1})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto (\mu n)^{-1}$
  - **Non-strongly convex:**  $O(n^{-1/2})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto n^{-1/2}$

# Convex stochastic approximation

## Existing work

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
  - **Strongly convex:**  $O((\mu n)^{-1})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto (\mu n)^{-1}$
  - **Non-strongly convex:**  $O(n^{-1/2})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto n^{-1/2}$
- **Many contributions in optimization and online learning:** Bottou and Le Cun (2005); Bottou and Bousquet (2008); Hazan et al. (2007); Shalev-Shwartz and Srebro (2008); Shalev-Shwartz et al. (2007, 2009); Xiao (2010); Duchi and Singer (2009); Nesterov and Vial (2008); Nemirovski et al. (2009)

# Convex stochastic approximation

## Existing work

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
  - **Strongly convex:**  $O((\mu n)^{-1})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto (\mu n)^{-1}$
  - **Non-strongly convex:**  $O(n^{-1/2})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto n^{-1/2}$
- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
  - All step sizes  $\gamma_n = Cn^{-\alpha}$  with  $\alpha \in (1/2, 1)$  lead to  $O(n^{-1})$  for **smooth** strongly convex problems



# Convex stochastic approximation

## Existing work

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
  - **Strongly convex:**  $O((\mu n)^{-1})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto (\mu n)^{-1}$
  - **Non-strongly convex:**  $O(n^{-1/2})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto n^{-1/2}$
- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
  - All step sizes  $\gamma_n = Cn^{-\alpha}$  with  $\alpha \in (1/2, 1)$  lead to  $O(n^{-1})$  for **smooth** strongly convex problems
- **Non-asymptotic analysis for smooth problems?**

# Smoothness/convexity assumptions

- Iteration:  $\theta_n = \theta_{n-1} - \gamma_n f'_n(\theta_{n-1})$ 
  - Polyak-Ruppert averaging:  $\bar{\theta}_n = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k$
- **Smoothness of  $f_n$** : For each  $n \geq 1$ , the function  $f_n$  is a.s. convex, differentiable with  $L$ -Lipschitz-continuous gradient  $f'_n$ :
  - Smooth loss and bounded data
- **Strong convexity of  $f$** : The function  $f$  is strongly convex with respect to the norm  $\|\cdot\|$ , with convexity constant  $\mu > 0$ :
  - Invertible population covariance matrix
  - or regularization by  $\frac{\mu}{2} \|\theta\|^2$

# Summary of new results (Bach and Moulines, 2011)

- Stochastic gradient descent with learning rate  $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old:  $O(n^{-1})$  rate achieved **without** averaging for  $\alpha = 1$
  - New:  $O(n^{-1})$  rate achieved **with** averaging for  $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants
  - Forgetting of initial conditions
  - Robustness to the choice of  $C$

# Summary of new results (Bach and Moulines, 2011)

- Stochastic gradient descent with learning rate  $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old:  $O(n^{-1})$  rate achieved **without** averaging for  $\alpha = 1$
  - New:  $O(n^{-1})$  rate achieved **with** averaging for  $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants
  - Forgetting of initial conditions
  - Robustness to the choice of  $C$
- **Convergence rates** for  $\mathbb{E}\|\theta_n - \theta_*\|^2$  and  $\mathbb{E}\|\bar{\theta}_n - \theta_*\|^2$ 
  - no averaging:  $O\left(\frac{\sigma^2 \gamma_n}{\mu}\right) + O(e^{-\mu n \gamma_n})\|\theta_0 - \theta_*\|^2$
  - averaging:  $\frac{\text{tr } H(\theta_*)^{-1}}{n} + \mu^{-1}O(n^{-2\alpha} + n^{-2+\alpha}) + O\left(\frac{\|\theta_0 - \theta_*\|^2}{\mu^2 n^2}\right)$

# Classical proof sketch (no averaging) - I

$$\begin{aligned}
\|\theta_n - \theta_*\|_2^2 &= \|\theta_{n-1} - \gamma_n f'_n(\theta_{n-1}) - \theta_*\|_2^2 \\
&= \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) + \gamma_n^2 \|f'_n(\theta_{n-1})\|_2^2 \\
&\leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \\
&\quad + 2\gamma_n^2 \|f'_n(\theta_*)\|_2^2 + 2\gamma_n^2 \|f'_n(\theta_{n-1}) - f'_n(\theta_*)\|_2^2 \\
&\leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \\
&\quad + 2\gamma_n^2 \|f'_n(\theta_*)\|_2^2 + 2\gamma_n^2 L [f'_n(\theta_{n-1}) - f'_n(\theta_*)]^\top (\theta_{n-1} - \theta_*) \\
\mathbb{E}[\|\theta_n - \theta_*\|_2^2 | \mathcal{F}_{n-1}] &\leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'(\theta_{n-1}) \\
&\quad + 2\gamma_n^2 \mathbb{E} \|f'_n(\theta_*)\|_2^2 + 2\gamma_n^2 L [f'(\theta_{n-1}) - 0]^\top (\theta_{n-1} - \theta_*) \\
&\leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (1 - \gamma_n L) (\theta_{n-1} - \theta_*)^\top f'(\theta_{n-1}) + 2\gamma_n^2 \sigma^2 \\
&\leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (1 - \gamma_n L) \frac{1}{2} \mu \|\theta_{n-1} - \theta_*\|_2^2 + 2\gamma_n^2 \sigma^2 \\
&= [1 - \mu \gamma_n (1 - \gamma_n L)] \|\theta_{n-1} - \theta_*\|_2^2 + 2\gamma_n^2 \sigma^2 \\
\mathbb{E}[\|\theta_n - \theta_*\|_2^2] &\leq [1 - \mu \gamma_n (1 - \gamma_n L)] \mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] + 2\gamma_n^2 \sigma^2
\end{aligned}$$

# Classical proof sketch (no averaging) - II

- **Main bound**

$$\begin{aligned}\mathbb{E}[\|\theta_n - \theta_*\|_2^2] &\leq [1 - \mu\gamma_n(1 - \gamma_n L)] \mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] + 2\gamma_n^2 \sigma^2 \\ &\leq [1 - \mu\gamma_n/2] \mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] + 2\gamma_n^2 \sigma^2 \text{ if } \gamma_n L \leq 1/2\end{aligned}$$

- **Classical results from stochastic approximation** (Kushner and Yin, 2003):  $\mathbb{E}[\|\theta_n - \theta_*\|_2^2]$  is smaller than

$$\begin{aligned}&\leq \prod_{i=1}^n [1 - \mu\gamma_i/2] \mathbb{E}[\|\theta_0 - \theta_*\|_2^2] + \sum_{k=1}^n \prod_{i=k+1}^n [1 - \mu\gamma_i/2] 2\gamma_k^2 \sigma^2 \\ &\leq \exp\left[-\frac{\mu}{2} \sum_{i=1}^n \gamma_i\right] \mathbb{E}[\|\theta_0 - \theta_*\|_2^2] + \sum_{k=1}^n \prod_{i=k+1}^n [1 - \mu\gamma_i/2] 2\gamma_k^2 \sigma^2\end{aligned}$$

## Decomposition of the noise term

- Assume  $(\gamma_n)$  is decreasing and less than  $1/\mu$ ; then for any  $m \in \{1, \dots, n\}$ , we may split the following sum as follows:

$$\begin{aligned}
 \sum_{k=1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 &= \sum_{k=1}^m \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 + \sum_{k=m+1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 \\
 &\leq \prod_{i=m+1}^n (1 - \mu\gamma_i) \sum_{k=1}^m \gamma_k^2 + \gamma_m \sum_{k=m+1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k \\
 &\leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^m \gamma_k^2 + \frac{\gamma_m}{\mu} \sum_{k=m+1}^n \left[ \prod_{i=k+1}^n (1 - \mu\gamma_i) - \prod_{i=k}^n (1 - \mu\gamma_i) \right] \\
 &\leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^m \gamma_k^2 + \frac{\gamma_m}{\mu} \left[ 1 - \prod_{i=m+1}^n (1 - \mu\gamma_i) \right] \\
 &\leq \exp\left(-\mu \sum_{i=m+1}^n \gamma_i\right) \sum_{k=1}^n \gamma_k^2 + \frac{\gamma_m}{\mu}, \text{ with e.g. } m = n/2
 \end{aligned}$$

# Decomposition of the noise term

$$\sum_{k=1}^n \prod_{i=k+1}^n (1 - \mu\gamma_i) \gamma_k^2 \leq \exp \left( -\mu \sum_{i=m+1}^n \gamma_i \right) \sum_{k=1}^n \gamma_k^2 + \frac{\gamma_m}{\mu}$$

- Require  $\gamma_n$  to tend to zero (vanishing decaying step-size)
  - May not need  $\sum_n \gamma_n^2 < \infty$  for convergence in quadratic mean
- Examples:  $\boxed{\gamma_n = C/n^\alpha}$ 
  - $\alpha = 1$ ,  $\sum_{i=1}^n \frac{1}{i} = \log(n) + \text{cst} + O(1/n)$
  - $\alpha > 1$ ,  $\sum_{i=1}^n \frac{1}{i^\alpha} = \text{cst} + O(1/n^{\alpha-1})$
  - $\alpha \in (0, 1)$ ,  $\sum_{i=1}^n \frac{1}{i^\alpha} = \text{cst} \times n^{1-\alpha} + O(1)$
  - Proof using relationship with integrals
  - Consequences: **need  $\alpha \in (0, 1)$**



# Proof sketch (averaging)

- From Polyak and Juditsky (1992):

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

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

$$\Leftrightarrow f'_n(\theta_*) + f''_n(\theta_*)(\theta_{n-1} - \theta_*) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n) + O(\|\theta_{n-1} - \theta_*\|^2)$$

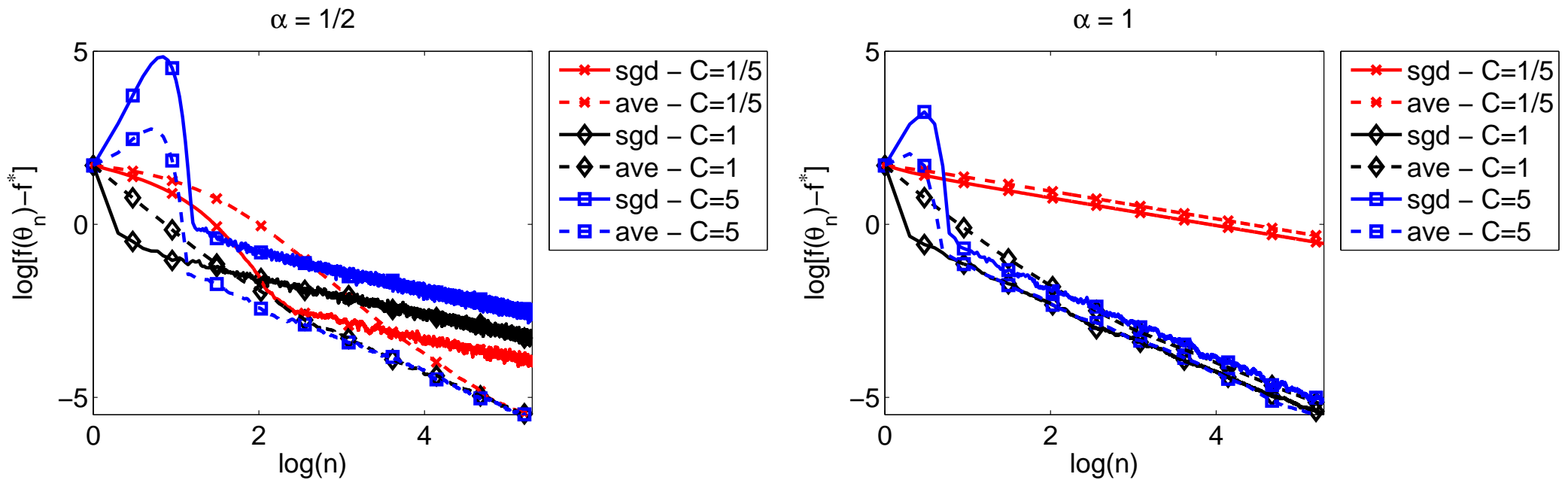
$$\Leftrightarrow f'_n(\theta_*) + f''(\theta_*)(\theta_{n-1} - \theta_*) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n) + O(\|\theta_{n-1} - \theta_*\|^2) \\ + O(\|\theta_{n-1} - \theta_*\|)\varepsilon_n$$

$$\Leftrightarrow \theta_{n-1} - \theta_* = -f''(\theta_*)^{-1} f'_n(\theta_*) + \frac{1}{\gamma_n} f''(\theta_*)^{-1}(\theta_{n-1} - \theta_n) \\ + O(\|\theta_{n-1} - \theta_*\|^2) + O(\|\theta_{n-1} - \theta_*\|)\varepsilon_n$$

- Averaging to cancel the term  $\frac{1}{\gamma_n} f''(\theta_*)^{-1}(\theta_{n-1} - \theta_n)$

# Robustness to wrong constants for $\gamma_n = Cn^{-\alpha}$

- $f(\theta) = \frac{1}{2}|\theta|^2$  with i.i.d. Gaussian noise ( $d = 1$ )
- Left:  $\alpha = 1/2$
- Right:  $\alpha = 1$



- See also <http://leon.bottou.org/projects/sgd>

# Summary of new results (Bach and Moulines, 2011)

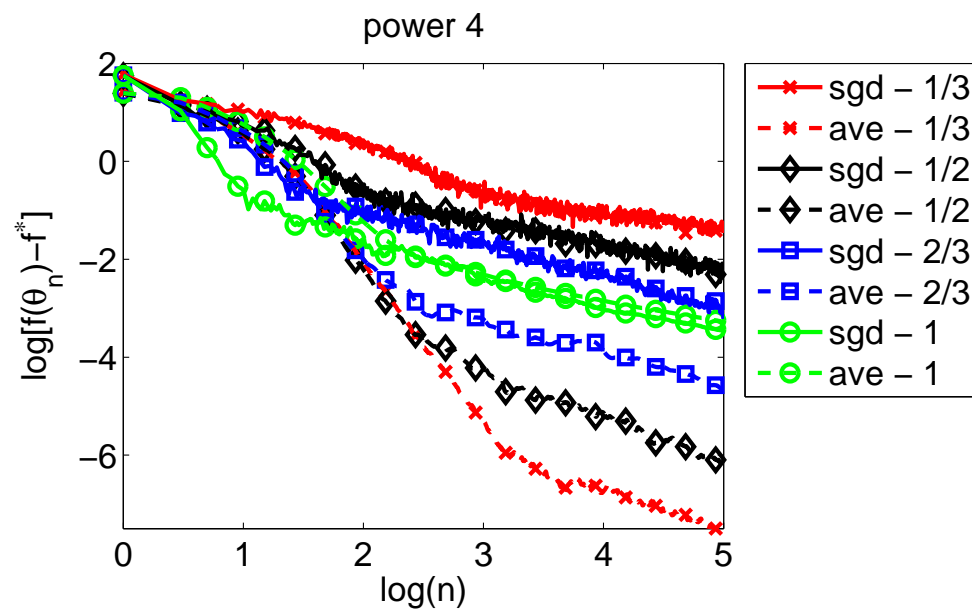
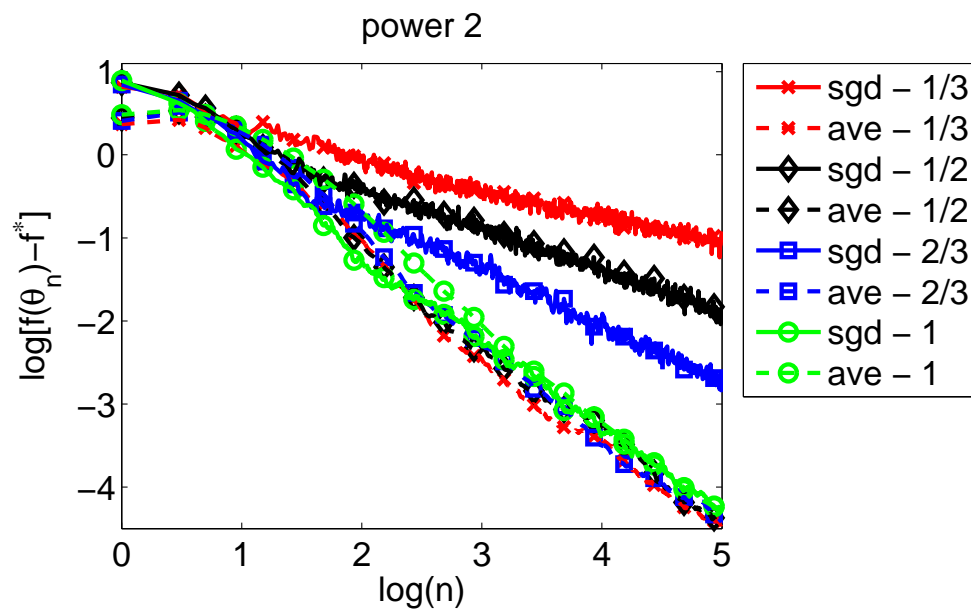
- Stochastic gradient descent with learning rate  $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old:  $O(n^{-1})$  rate achieved **without** averaging for  $\alpha = 1$
  - New:  $O(n^{-1})$  rate achieved **with** averaging for  $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants

# Summary of new results (Bach and Moulines, 2011)

- Stochastic gradient descent with learning rate  $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old:  $O(n^{-1})$  rate achieved **without** averaging for  $\alpha = 1$
  - New:  $O(n^{-1})$  rate achieved **with** averaging for  $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants
- **Non-strongly convex smooth objective functions**
  - Old:  $O(n^{-1/2})$  rate achieved **with** averaging for  $\alpha = 1/2$
  - New:  $O(\max\{n^{1/2-3\alpha/2}, n^{-\alpha/2}, n^{\alpha-1}\})$  rate achieved **without** averaging for  $\alpha \in [1/3, 1]$
- **Take-home message**
  - Use  $\alpha = 1/2$  with averaging to be adaptive to strong convexity

# Robustness to lack of strong convexity

- Left:  $f(\theta) = |\theta|^2$  between  $-1$  and  $1$
- Right:  $f(\theta) = |\theta|^4$  between  $-1$  and  $1$
- affine outside of  $[-1, 1]$ , continuously differentiable.



# Convex stochastic approximation

## Existing work

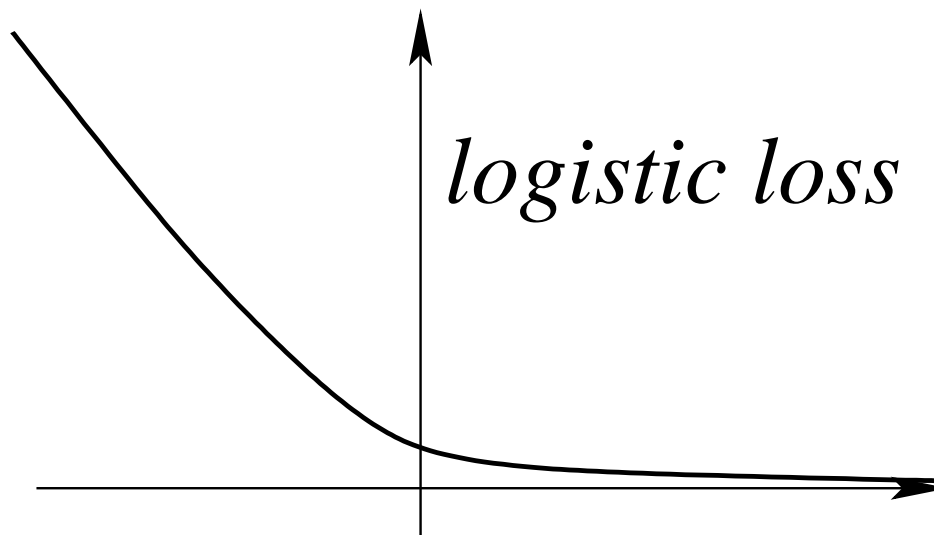
- **Known global minimax rates of convergence for non-smooth problems** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
  - **Strongly convex:**  $O((\mu n)^{-1})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto (\mu n)^{-1}$
  - **Non-strongly convex:**  $O(n^{-1/2})$   
Attained by averaged stochastic gradient descent with  $\gamma_n \propto n^{-1/2}$
- **Asymptotic analysis of averaging** (Polyak and Juditsky, 1992; Ruppert, 1988)
  - All step sizes  $\gamma_n = Cn^{-\alpha}$  with  $\alpha \in (1/2, 1)$  lead to  $O(n^{-1})$  for **smooth** strongly convex problems
- **A single adaptive algorithm for smooth problems with convergence rate  $O(\min\{1/\mu n, 1/\sqrt{n}\})$  in all situations?**

# Adaptive algorithm for logistic regression

- **Logistic regression:**  $(\Phi(x_n), y_n) \in \mathbb{R}^d \times \{-1, 1\}$ 
  - Single data point:  $f_n(\theta) = \log(1 + \exp(-y_n \theta^\top \Phi(x_n)))$
  - Generalization error:  $f(\theta) = \mathbb{E} f_n(\theta)$

# Adaptive algorithm for logistic regression

- **Logistic regression:**  $(\Phi(x_n), y_n) \in \mathbb{R}^d \times \{-1, 1\}$ 
  - Single data point:  $f_n(\theta) = \log(1 + \exp(-y_n \theta^\top \Phi(x_n)))$
  - Generalization error:  $f(\theta) = \mathbb{E} f_n(\theta)$
- **Cannot be strongly convex**  $\Rightarrow$  **local** strong convexity
  - unless restricted to  $|\theta^\top \Phi(x_n)| \leq M$  (with constants  $e^M$  - *proof*)
  - $\mu =$  lowest eigenvalue of the Hessian at the optimum  $f''(\theta_*)$





# Adaptive algorithm for logistic regression

- **Logistic regression:**  $(\Phi(x_n), y_n) \in \mathbb{R}^d \times \{-1, 1\}$ 
  - Single data point:  $f_n(\theta) = \log(1 + \exp(-y_n \theta^\top \Phi(x_n)))$
  - Generalization error:  $f(\theta) = \mathbb{E} f_n(\theta)$
- **Cannot be strongly convex**  $\Rightarrow$  **local** strong convexity
  - unless restricted to  $|\theta^\top \Phi(x_n)| \leq M$  (with constants  $e^M$  - *proof*)
  - $\mu$  = lowest eigenvalue of the Hessian at the optimum  $f''(\theta_*)$
- **$n$  steps of averaged SGD with constant step-size  $1/(2R^2\sqrt{n})$** 
  - with  $R$  = radius of data (Bach, 2013):
$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \min \left\{ \frac{1}{\sqrt{n}}, \frac{R^2}{n\mu} \right\} (15 + 5R\|\theta_0 - \theta_*\|)^4$$
  - Proof based on self-concordance (Nesterov and Nemirovski, 1994)

# Self-concordance - I

- Usual definition for convex  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ :  $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$ 
  - Affine invariant
  - Extendable to all convex functions on  $\mathbb{R}^d$  by looking at rays
  - Used for the sharp proof of quadratic convergence of Newton method (Nesterov and Nemirovski, 1994)
- Generalized notion:  $|\varphi'''(t)| \leq \varphi''(t)$ 
  - Applicable to logistic regression (with extensions)
  - $\varphi(t) = \log(1 + e^{-t})$ ,  $\varphi'(t) = (1 + e^t)^{-1}$ , etc...

# Self-concordance - I

- Usual definition for convex  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ :  $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$ 
  - Affine invariant
  - Extendable to all convex functions on  $\mathbb{R}^d$  by looking at rays
  - Used for the sharp proof of quadratic convergence of Newton method (Nesterov and Nemirovski, 1994)
- Generalized notion:  $|\varphi'''(t)| \leq \varphi''(t)$ 
  - Applicable to logistic regression (with extensions)
  - If features bounded by  $R$ ,  $h : t \mapsto f[\theta_1 + t(\theta_2 - \theta_1)]$  satisfies:  
 $\forall t \in \mathbb{R}, |h'''(t)| \leq R\|\theta_1 - \theta_2\|h''(t)$

# Self-concordance - I

- Usual definition for convex  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ :  $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$ 
  - Affine invariant
  - Extendable to all convex functions on  $\mathbb{R}^d$  by looking at rays
  - Used for the sharp proof of quadratic convergence of Newton method (Nesterov and Nemirovski, 1994)
- Generalized notion:  $|\varphi'''(t)| \leq \varphi''(t)$ 
  - Applicable to logistic regression (with extensions)
  - If features bounded by  $R$ ,  $h : t \mapsto f[\theta_1 + t(\theta_2 - \theta_1)]$  satisfies:  
 $\forall t \in \mathbb{R}, |h'''(t)| \leq R\|\theta_1 - \theta_2\|h''(t)$
- **Important properties**
  - Allows global Taylor expansions
  - Relates expansions of derivatives of different orders

## Global Taylor expansions

- **Lemma:** If  $\forall t \in \mathbb{R}, |g'''(t)| \leq Sg''(t)$ , for  $S \geq 0$ . Then,  $\forall t \geq 0$ :

$$\frac{g''(0)}{S^2}(e^{-St} + St - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{S^2}(e^{St} - St - 1)$$

# Global Taylor expansions

- **Lemma:** If  $\forall t \in \mathbb{R}, |g'''(t)| \leq Sg''(t)$ , for  $S \geq 0$ . Then,  $\forall t \geq 0$ :

$$\frac{g''(0)}{S^2}(e^{-St} + St - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{S^2}(e^{St} - St - 1)$$

- **Proof:** Let us first assume that  $g''(t)$  is strictly positive for all  $t \in \mathbb{R}$ . We have, for all  $t \geq 0$ :  $-S \leq \frac{d \log g''(t)}{dt} \leq S$ . Then, by integrating once between 0 and  $t$ , taking exponentials, and then integrating twice:

$$-St \leq \log g''(t) - \log g''(0) \leq St,$$

$$g''(0)e^{-St} \leq g''(t) \leq g''(0)e^{St}, \quad (1)$$

$$g''(0)S^{-1}(1 - e^{-St}) \leq g'(t) - g'(0) \leq g''(0)S^{-1}(e^{St} - 1),$$

$$g(t) \geq g(0) + g'(0)t + g''(0)S^{-2}(e^{-St} + St - 1), \quad (2)$$

$$g(t) \leq g(0) + g'(0)t + g''(0)S^{-2}(e^{St} - St - 1), \quad (3)$$

which leads to the desired result (simple reasoning for strict positivity of  $g''$ )

## Relating Taylor expansions of different orders

- **Lemma:** If  $h : t \mapsto f[\theta_1 + t(\theta_2 - \theta_1)]$  satisfies:  $\forall t \in \mathbb{R}, |h'''(t)| \leq R\|\theta_1 - \theta_2\|h''(t)$ . We have, for all  $\theta_1, \theta_2 \in \mathbb{R}^d$ :  
$$\|f'(\theta_1) - f'(\theta_2) - f''(\theta_2)(\theta_2 - \theta_1)\| \leq R[f(\theta_1) - f(\theta_2) - \langle f'(\theta_2), \theta_2 - \theta_1 \rangle]$$

## Relating Taylor expansions of different orders

- **Lemma:** If  $h : t \mapsto f[\theta_1 + t(\theta_2 - \theta_1)]$  satisfies:  $\forall t \in \mathbb{R}, |h'''(t)| \leq R\|\theta_1 - \theta_2\|h''(t)$ . We have, for all  $\theta_1, \theta_2 \in \mathbb{R}^d$ :

$$\|f'(\theta_1) - f'(\theta_2) - f''(\theta_2)(\theta_2 - \theta_1)\| \leq R[f(\theta_1) - f(\theta_2) - \langle f'(\theta_2), \theta_2 - \theta_1 \rangle]$$

- **Proof:** For  $\|z\| = 1$ , let  $\varphi(t) = \langle z, f'(\theta_2 + t(\theta_1 - \theta_2)) - f'(\theta_2) - tf''(\theta_2)(\theta_2 - \theta_1) \rangle$  and  $\psi(t) = R[f(\theta_2 + t(\theta_1 - \theta_2)) - f(\theta_2) - t\langle f'(\theta_2), \theta_2 - \theta_1 \rangle]$ . Then  $\varphi(0) = \psi(0) = 0$ , and:

$$\varphi'(t) = \langle z, f''(\theta_2 + t(\theta_1 - \theta_2)) - f''(\theta_2), \theta_1 - \theta_2 \rangle$$

$$\varphi''(t) = f'''(\theta_2 + t(\theta_1 - \theta_2))[z, \theta_1 - \theta_2, \theta_1 - \theta_2]$$

$$\leq R\|z\|_2 f''(\theta_2 + t(\theta_1 - \theta_2))[\theta_1 - \theta_2, \theta_1 - \theta_2], \text{ using App. A of Bach (2010)}$$

$$= R\langle \theta_2 - \theta_1, f''(\theta_2 + t(\theta_1 - \theta_2))(\theta_1 - \theta_2) \rangle$$

$$\psi'(t) = R\langle f'(\theta_2 + t(\theta_1 - \theta_2)) - f'(\theta_2), \theta_1 - \theta_2 \rangle$$

$$\psi''(t) = R\langle \theta_2 - \theta_1, f''(\theta_2 + t(\theta_1 - \theta_2))(\theta_1 - \theta_2) \rangle,$$

Thus  $\varphi'(0) = \psi'(0) = 0$  and  $\varphi''(t) \leq \psi''(t)$ , leading to  $\varphi(1) \leq \psi(1)$  by integrating twice, which leads to the desired result by maximizing with respect to  $z$ .



# Adaptive algorithm for logistic regression

## Proof sketch

- Step 1: use existing result  $f(\bar{\theta}_n) - f(\theta_*) + \frac{R^2}{\sqrt{n}} \|\theta_0 - \theta_*\|_2^2 = O(1/\sqrt{n})$
- Step 2a:  $f'_n(\theta_{n-1}) = \frac{1}{\gamma}(\theta_{n-1} - \theta_n) \Rightarrow \frac{1}{n} \sum_{k=1}^n f'_k(\theta_{k-1}) = \frac{1}{n\gamma}(\theta_0 - \theta_n)$
- Step 2b:  $\frac{1}{n} \sum_{k=1}^n f'(\theta_{k-1}) = \frac{1}{n} \sum_{k=1}^n [f'(\theta_{k-1}) - f'_k(\theta_{k-1})] + \frac{1}{\gamma n}(\theta_0 - \theta_*) + \frac{1}{\gamma n}(\theta_* - \theta_n) = O(1/\sqrt{n})$
- Step 3:  $\left\| f'\left(\frac{1}{n} \sum_{k=1}^n \theta_{k-1}\right) - \frac{1}{n} \sum_{k=1}^n f'(\theta_{k-1}) \right\|_2 = O(f(\bar{\theta}_n) - f(\theta_*)) = O(1/\sqrt{n})$  using self-concordance
- Step 4a: if  $f$   $\mu$ -strongly convex,  $f(\bar{\theta}_n) - f(\theta_*) \leq \frac{1}{2\mu} \|f'(\bar{\theta}_n)\|_2^2$
- Step 4b: if  $f$  self-concordant, “locally true” with  $\mu = \lambda_{\min}(f''(\theta_*))$

# Adaptive algorithm for logistic regression

- **Logistic regression:**  $(\Phi(x_n), y_n) \in \mathbb{R}^d \times \{-1, 1\}$ 
  - Single data point:  $f_n(\theta) = \log(1 + \exp(-y_n \theta^\top \Phi(x_n)))$
  - Generalization error:  $f(\theta) = \mathbb{E} f_n(\theta)$
- **Cannot be strongly convex**  $\Rightarrow$  **local** strong convexity
  - unless restricted to  $|\theta^\top \Phi(x_n)| \leq M$  (and with constants  $e^M$ )
  - $\mu =$  lowest eigenvalue of the Hessian at the optimum  $f''(\theta_*)$
- **$n$  steps of averaged SGD with constant step-size  $1/(2R^2\sqrt{n})$** 
  - with  $R =$  radius of data (Bach, 2013):
$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \min \left\{ \frac{1}{\sqrt{n}}, \frac{R^2}{n\mu} \right\} (15 + 5R\|\theta_0 - \theta_*\|)^4$$
  - Proof based on self-concordance (Nesterov and Nemirovski, 1994)

# Adaptive algorithm for logistic regression

- **Logistic regression:**  $(\Phi(x_n), y_n) \in \mathbb{R}^d \times \{-1, 1\}$ 
  - Single data point:  $f_n(\theta) = \log(1 + \exp(-y_n \theta^\top \Phi(x_n)))$
  - Generalization error:  $f(\theta) = \mathbb{E} f_n(\theta)$
- **Cannot be strongly convex**  $\Rightarrow$  **local** strong convexity
  - unless restricted to  $|\theta^\top \Phi(x_n)| \leq M$  (and with constants  $e^M$ )
  - $\mu =$  lowest eigenvalue of the Hessian at the optimum  $f''(\theta_*)$
- **$n$  steps of averaged SGD with constant step-size  $1/(2R^2\sqrt{n})$** 
  - with  $R =$  radius of data (Bach, 2013):

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \min \left\{ \frac{1}{\sqrt{n}}, \frac{R^2}{n\mu} \right\} (15 + 5R\|\theta_0 - \theta_*\|)^4$$

- **A single adaptive algorithm for smooth problems with convergence rate  $O(1/n)$  in all situations?**

# Least-mean-square algorithm

- **Least-squares:**  $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \Phi(x_n), \theta \rangle)^2]$  with  $\theta \in \mathbb{R}^d$ 
  - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
  - usually studied without averaging and decreasing step-sizes
  - with strong convexity assumption  $\mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)] = H \succcurlyeq \mu \cdot \text{Id}$

# Least-mean-square algorithm

- **Least-squares:**  $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \Phi(x_n), \theta \rangle)^2]$  with  $\theta \in \mathbb{R}^d$ 
  - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
  - usually studied without averaging and decreasing step-sizes
  - with strong convexity assumption  $\mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)] = H \succcurlyeq \mu \cdot \text{Id}$
- **New analysis for averaging and constant step-size**  $\gamma = 1/(4R^2)$ 
  - Assume  $\|\Phi(x_n)\| \leq R$  and  $|y_n - \langle \Phi(x_n), \theta_* \rangle| \leq \sigma$  almost surely
  - **No assumption regarding lowest eigenvalues of  $H$**
  - Main result: 
$$\mathbb{E}f(\bar{\theta}_{n-1}) - f(\theta_*) \leq \frac{4\sigma^2 d}{n} + \frac{4R^2 \|\theta_0 - \theta_*\|^2}{n}$$
- **Matches statistical lower bound** (Tsybakov, 2003)
  - Non-asymptotic robust version of Györfi and Walk (1996)

# Least-squares - Proof technique - I

- LMS recursion:

$$\theta_n - \theta_* = [I - \gamma \Phi(x_n) \otimes \Phi(x_n)] (\theta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n)$$

- Simplified LMS recursion: with  $H = \mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)]$

$$\theta_n - \theta_* = [I - \gamma H] (\theta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n)$$

- Direct proof technique of Polyak and Juditsky (1992), e.g.,

$$\theta_n - \theta_* = [I - \gamma H]^n (\theta_0 - \theta_*) + \gamma \sum_{k=1}^n [I - \gamma H]^{n-k} \varepsilon_k \Phi(x_k)$$

- Infinite expansion of Aguech, Moulines, and Priouret (2000) in powers of  $\gamma$

# Least-squares - Proof technique - II

- Explicit expansion of  $\bar{\theta}_n$ :

$$\theta_n - \theta_* = [I - \gamma H]^n (\theta_0 - \theta_*) + \gamma \sum_{k=1}^n [I - \gamma H]^{n-k} \varepsilon_k \Phi(x_k)$$

$$\begin{aligned} \bar{\theta}_n - \theta_* &= \frac{1}{n+1} \sum_{i=0}^n [I - \gamma H]^i (\theta_0 - \theta_*) + \frac{\gamma}{n+1} \sum_{i=0}^n \sum_{k=1}^i [I - \gamma H]^{i-k} \varepsilon_k \Phi(x_k) \\ &\approx \frac{1}{n} (\gamma H)^{-1} [I - (I - \gamma H)^n] (\theta_0 - \theta_*) + \frac{\gamma}{n} \sum_{k=0}^n (\gamma H)^{-1} \varepsilon_k \Phi(x_k) \end{aligned}$$

- Need to bound  $(\mathbb{E} \|H^{1/2}(\bar{\theta}_n - \theta_*)\|^2)^{1/2}$
- Using Minkowski inequality

## Least-squares - Proof technique - III

- Explicit expansion of  $\bar{\theta}_n$ :

$$\bar{\theta}_n - \theta_* \approx \frac{1}{n}(\gamma H)^{-1} [I - (I - \gamma H)^n] (\theta_0 - \theta_*) + \frac{\gamma}{n} \sum_{k=0}^n (\gamma H)^{-1} \varepsilon_k \Phi(x_k)$$

- **Bias - I:**  $(\gamma H)^{-1} [I - (I - \gamma H)^n] \asymp (\gamma H)^{-1}$  leading to

$$(\mathbb{E} \|H^{1/2}(\bar{\theta}_n - \theta_*)\|^2)^{1/2} \leq \frac{1}{\gamma n} \|H^{-1/2}(\theta_0 - \theta_*)\|$$

- **Bias - II:**  $(\gamma H)^{-1} [I - (I - \gamma H)^n] \asymp \sqrt{n}(\gamma H)^{-1/2}$  leading to

$$(\mathbb{E} \|H^{1/2}(\bar{\theta}_n - \theta_*)\|^2)^{1/2} \leq \frac{1}{\sqrt{\gamma n}} \|(\theta_0 - \theta_*)\|$$

- **Variance** (next slide)



# Least-squares - Proof technique - III

- Explicit expansion of  $\bar{\theta}_n$ :

$$\bar{\theta}_n - \theta_* \approx \frac{1}{n}(\gamma H)^{-1} [I - (I - \gamma H)^n] (\theta_0 - \theta_*) + \frac{\gamma}{n} \sum_{k=0}^n (\gamma H)^{-1} \varepsilon_k \Phi(x_k)$$

- **Variance** (next slide)

$$\begin{aligned} \mathbb{E} \| H^{1/2} (\bar{\theta}_n - \theta_*) \|^2 &= \frac{1}{n^2} \sum_{k=0}^n \mathbb{E} \varepsilon_k^2 \langle \Phi(x_k), H^{-1} \Phi(x_k) \rangle \\ &= \frac{1}{n} \sigma^2 d \end{aligned}$$

## Least-squares - Proof technique - IV

- Expansion of Aguech, Moulines, and Priouret (2000) in powers of  $\gamma$ 
  - LMS recursion:

$$\theta_n - \theta_* = [I - \gamma \Phi(x_n) \otimes \Phi(x_n)] (\theta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n)$$

- Simplified LMS recursion: with  $H = \mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)]$

$$\eta_n - \theta_* = [I - \gamma H] (\eta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n)$$

- Expansion of the difference:

$$\theta_n - \eta_n = [I - \gamma \Phi(x_n) \otimes \Phi(x_n)] (\theta_{n-1} - \eta_{n-1}) + \gamma [H - \Phi(x_n) \otimes \Phi(x_n)] (\eta_{n-1} - \theta_*)$$

## Least-squares - Proof technique - IV

- Expansion of Aguech, Moulines, and Priouret (2000) in powers of  $\gamma$ 
  - LMS recursion:

$$\theta_n - \theta_* = [I - \gamma \Phi(x_n) \otimes \Phi(x_n)] (\theta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n)$$

- Simplified LMS recursion: with  $H = \mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)]$

$$\eta_n - \theta_* = [I - \gamma H] (\eta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n)$$

- Expansion of the difference:

$$\theta_n - \eta_n = [I - \gamma \Phi(x_n) \otimes \Phi(x_n)] (\theta_{n-1} - \eta_{n-1}) + \gamma [H - \Phi(x_n) \otimes \Phi(x_n)] (\eta_{n-1} - \theta_*)$$

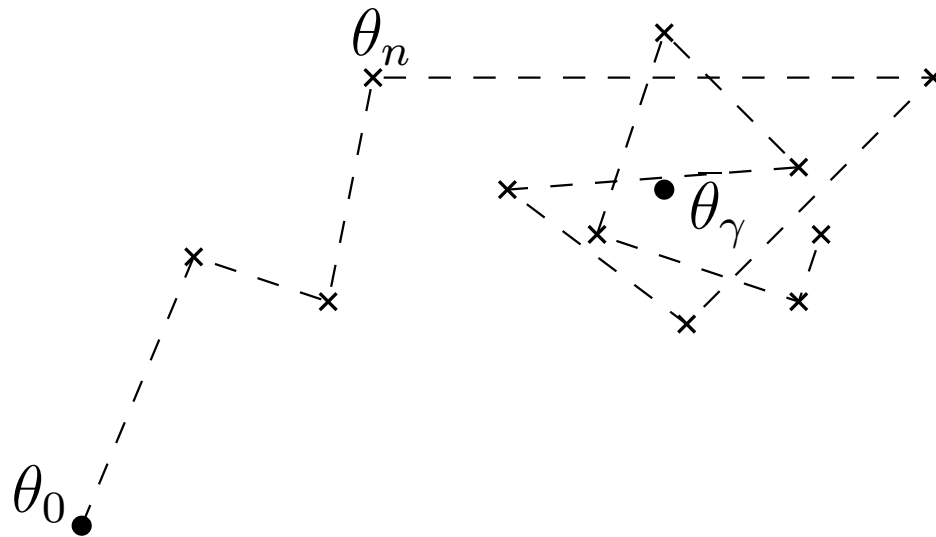
- New noise process
- May continue the expansion infinitely many times

# Markov chain interpretation of constant step sizes

- LMS recursion for  $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence  $(\theta_n)_n$  is a **homogeneous Markov chain**
  - convergence to a stationary distribution  $\pi_\gamma$
  - with expectation  $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$



# Markov chain interpretation of constant step sizes

- LMS recursion for  $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

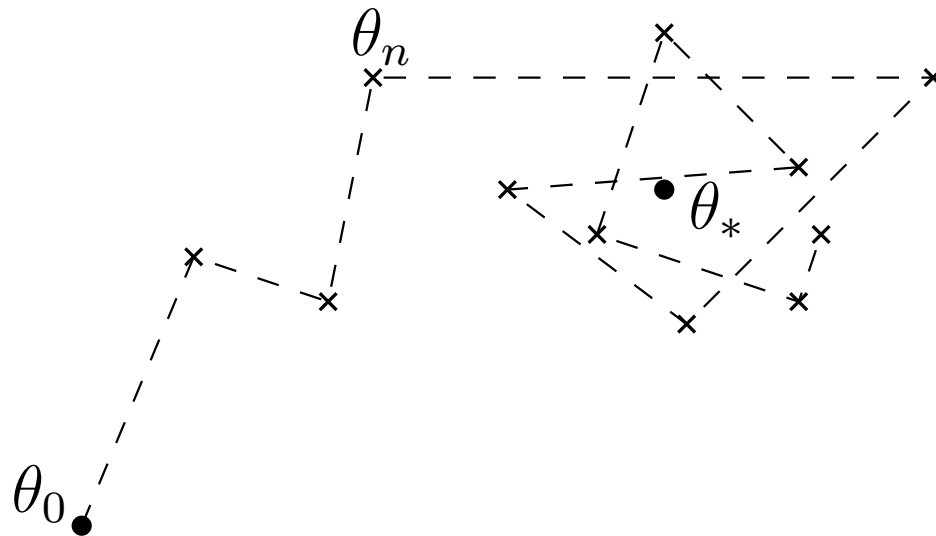
$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence  $(\theta_n)_n$  is a **homogeneous Markov chain**

– convergence to a stationary distribution  $\pi_\gamma$

– with expectation  $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

- **For least-squares,  $\bar{\theta}_\gamma = \theta_*$**



# Markov chain interpretation of constant step sizes

- LMS recursion for  $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

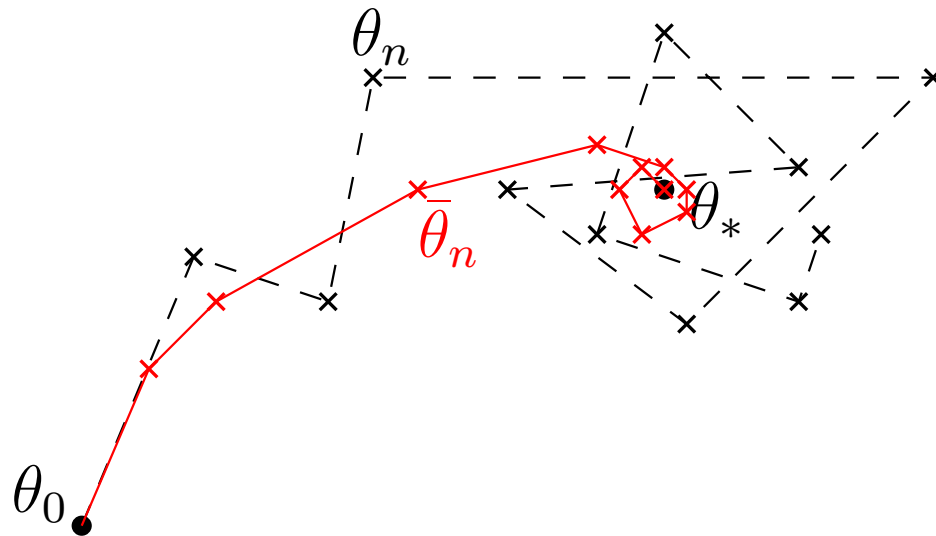
$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence  $(\theta_n)_n$  is a **homogeneous Markov chain**

– convergence to a stationary distribution  $\pi_\gamma$

– with expectation  $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

- **For least-squares,  $\bar{\theta}_\gamma = \theta_*$**



# Markov chain interpretation of constant step sizes

- LMS recursion for  $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$

$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n)$$

- The sequence  $(\theta_n)_n$  is a **homogeneous Markov chain**

- convergence to a stationary distribution  $\pi_\gamma$

- with expectation  $\bar{\theta}_\gamma \stackrel{\text{def}}{=} \int \theta \pi_\gamma(d\theta)$

- **For least-squares,  $\bar{\theta}_\gamma = \theta_*$**

- $\theta_n$  does not converge to  $\theta_*$  but oscillates around it

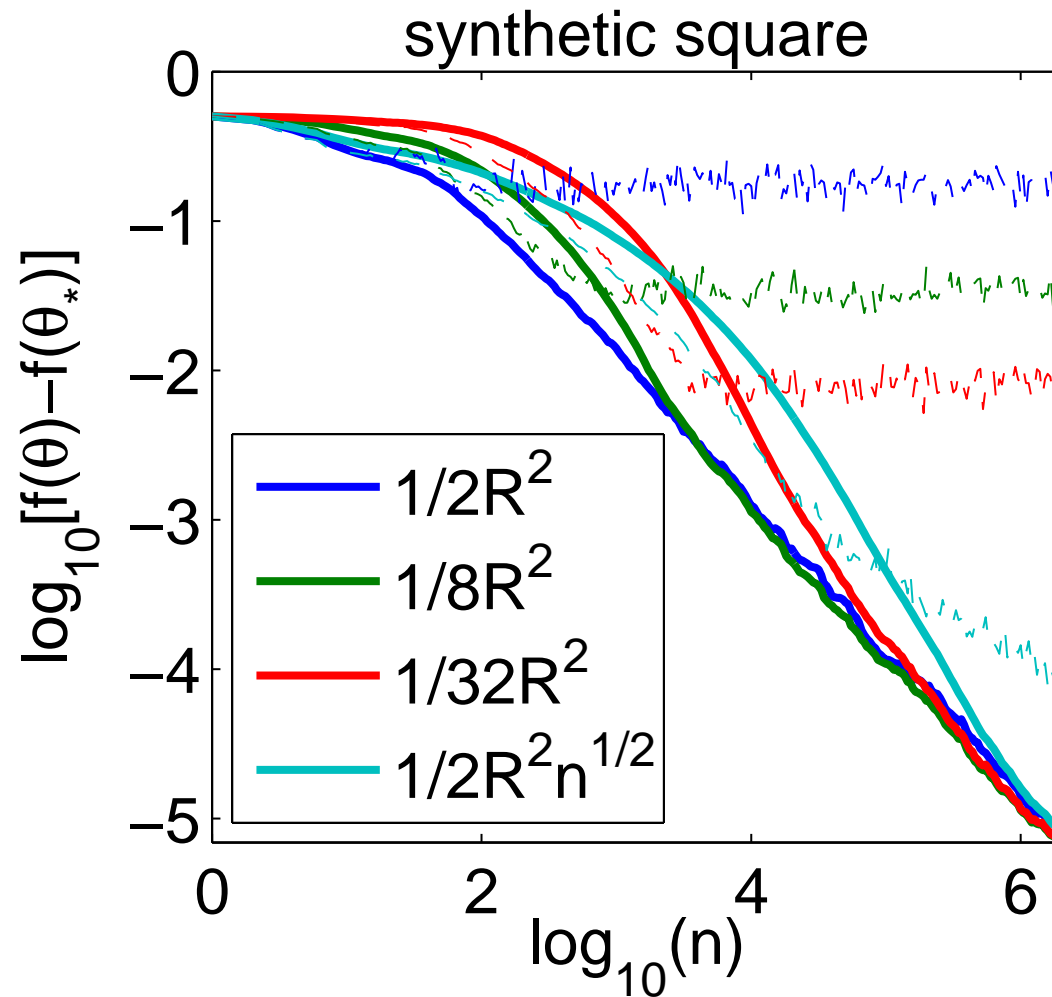
- oscillations of order  $\sqrt{\gamma}$

- **Ergodic theorem:**

- Averaged iterates converge to  $\bar{\theta}_\gamma = \theta_*$  at rate  $O(1/n)$

# Simulations - synthetic examples

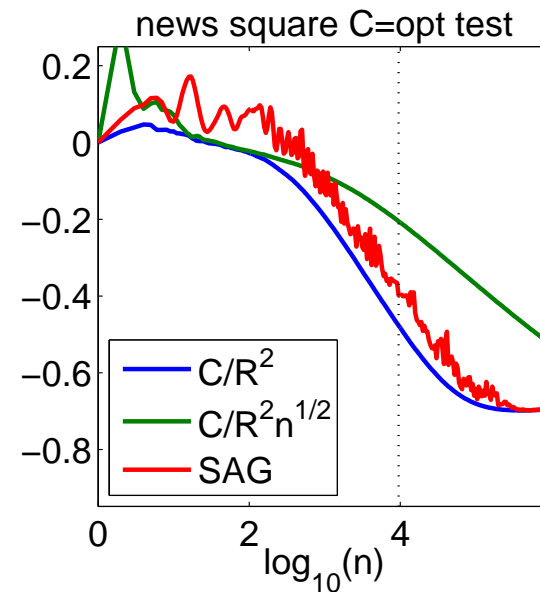
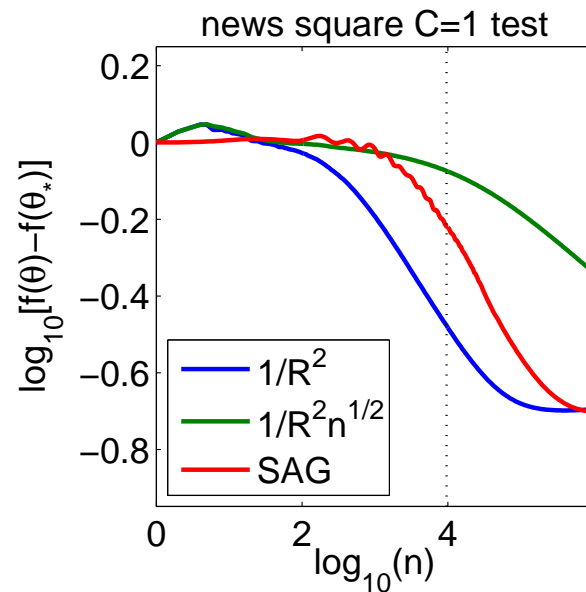
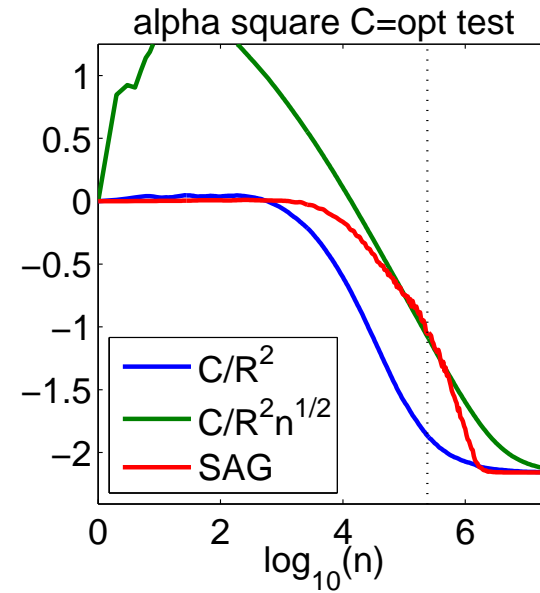
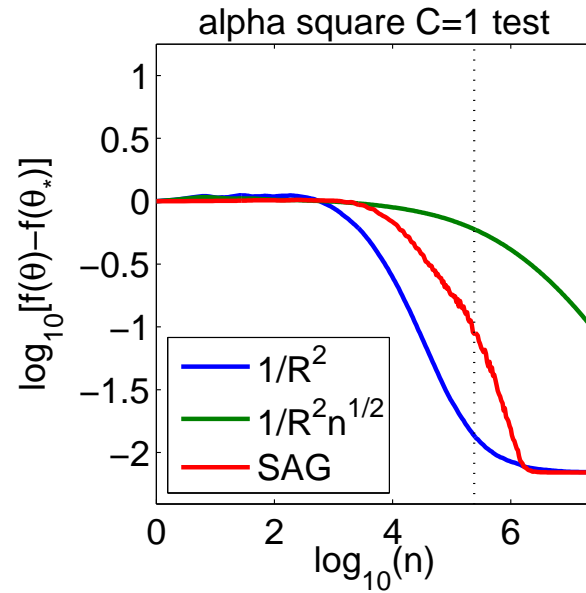
- Gaussian distributions -  $d = 20$





# Simulations - benchmarks

- *alpha* ( $d = 500, n = 500\ 000$ ), *news* ( $d = 1\ 300\ 000, n = 20\ 000$ )



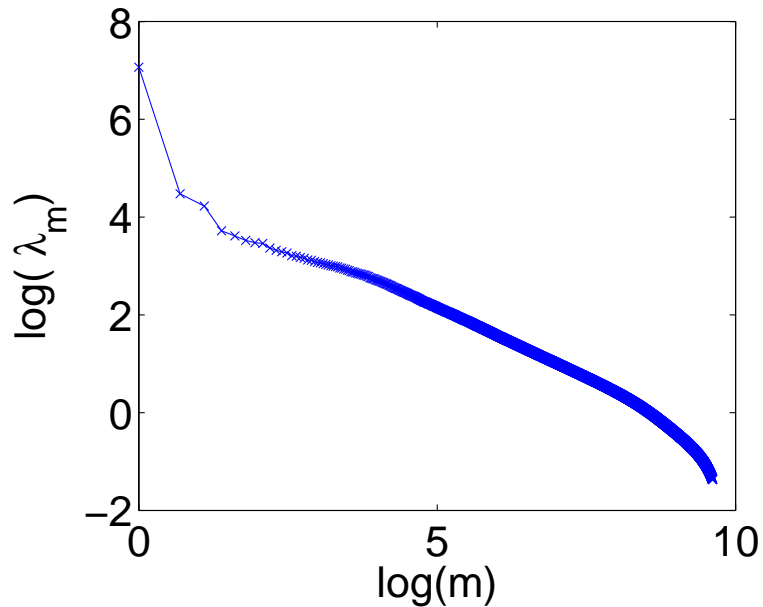
# Optimal bounds for least-squares?

- **Least-squares:** cannot beat  $\sigma^2 d/n$  (Tsybakov, 2003). Really?
  - What if  $d \gg n$ ?
- **Refined assumptions with adaptivity** (Dieuleveut and Bach, 2014)
  - Beyond strong convexity or lack thereof

# Finer assumptions (Dieuleveut and Bach, 2014)

- **Covariance eigenvalues**

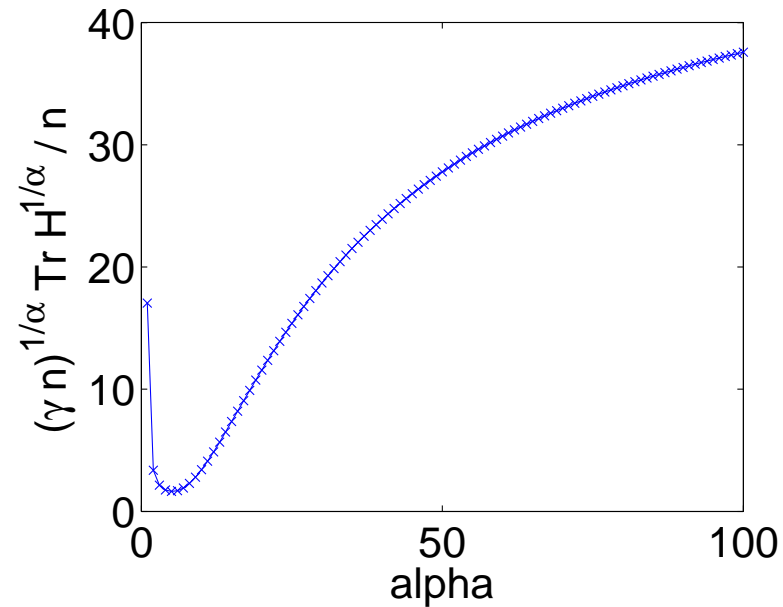
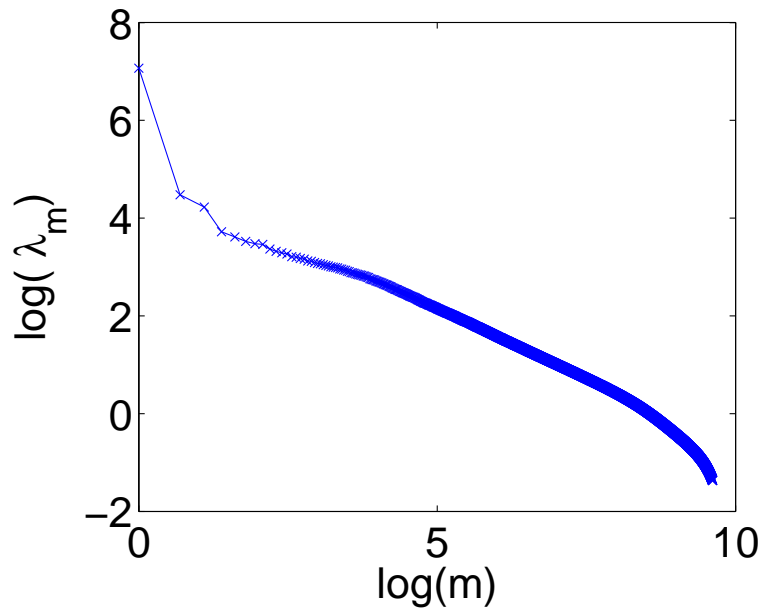
- Pessimistic assumption: all eigenvalues  $\lambda_m$  less than a constant
- Actual decay as  $\lambda_m = o(m^{-\alpha})$  with  $\text{tr } H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$  small



# Finer assumptions (Dieuleveut and Bach, 2014)

- Covariance eigenvalues

- Pessimistic assumption: all eigenvalues  $\lambda_m$  less than a constant
- Actual decay as  $\lambda_m = o(m^{-\alpha})$  with  $\text{tr } H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$  small
- New result: replace  $\frac{\sigma^2 d}{n}$  by  $\frac{\sigma^2 (\gamma n)^{1/\alpha} \text{tr } H^{1/\alpha}}{n}$



# Finer assumptions (Dieuleveut and Bach, 2014)

- **Covariance eigenvalues**

- Pessimistic assumption: all eigenvalues  $\lambda_m$  less than a constant
- Actual decay as  $\lambda_m = o(m^{-\alpha})$  with  $\text{tr } H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$  small
- New result: replace  $\frac{\sigma^2 d}{n}$  by  $\frac{\sigma^2 (\gamma n)^{1/\alpha} \text{tr } H^{1/\alpha}}{n}$

- **Optimal predictor**

- Pessimistic assumption:  $\|\theta_0 - \theta_*\|^2$  finite
- Finer assumption:  $\|H^{1/2-r}(\theta_0 - \theta_*)\|_2$  small
- Replace  $\frac{\|\theta_0 - \theta_*\|^2}{\gamma n}$  by  $\frac{4\|H^{1/2-r}(\theta_0 - \theta_*)\|_2}{\gamma^{2r} n^{2 \min\{r,1\}}}$

# Optimal bounds for least-squares?

- **Least-squares:** cannot beat  $\sigma^2 d/n$  (Tsybakov, 2003). Really?
  - What if  $d \gg n$ ?
- **Refined assumptions with adaptivity** (Dieuleveut and Bach, 2014)
  - Beyond strong convexity or lack thereof

$$f(\bar{\theta}_n) - f(\theta_*) \leq \frac{16\sigma^2 \operatorname{tr} H^{1/\alpha}}{n} (\gamma n)^{1/\alpha} + \frac{4 \|H^{1/2-r}(\theta_0 - \theta_*)\|_2}{\gamma^{2r} n^{2 \min\{r, 1\}}}$$

- Previous results:  $\alpha = +\infty$  and  $r = 1/2$
- Valid for all  $\alpha$  and  $r$
- Optimal step-size potentially decaying with  $n$
- Extension to non-parametric estimation (kernels) with optimal rates

# From least-squares to non-parametric estimation - I

- **Extension to Hilbert spaces:**  $\Phi(x), \theta \in \mathcal{H}$

$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n) \Phi(x_n)$$

- **If  $\theta_0 = 0$ ,  $\theta_n$  is a linear combination of  $\Phi(x_1), \dots, \Phi(x_n)$**

$$\theta_n = \sum_{k=1}^n \alpha_k \Phi(x_k) \quad \text{and} \quad \alpha_n = -\gamma \sum_{k=1}^{n-1} \alpha_k \langle \Phi(x_k), \Phi(x_n) \rangle + \gamma y_n$$

# From least-squares to non-parametric estimation - I

- **Extension to Hilbert spaces:**  $\Phi(x), \theta \in \mathcal{H}$

$$\theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n) \Phi(x_n)$$

- **If  $\theta_0 = 0$ ,  $\theta_n$  is a linear combination of  $\Phi(x_1), \dots, \Phi(x_n)$**

$$\theta_n = \sum_{k=1}^n \alpha_k \Phi(x_k) \quad \text{and} \quad \alpha_n = -\gamma \sum_{k=1}^{n-1} \alpha_k \langle \Phi(x_k), \Phi(x_n) \rangle + \gamma y_n$$

- **Kernel trick:**  $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$

- Reproducing kernel Hilbert spaces and non-parametric estimation
- See, e.g., Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004); Dieuleveut and Bach (2014)
- Still  $O(n^2)$



# From least-squares to non-parametric estimation - II

- **Simple example:** Sobolev space on  $\mathcal{X} = [0, 1]$ 
  - $\Phi(x) =$  weighted Fourier basis  $\Phi(x)_j = \varphi_j \cos(2j\pi x)$  (plus sine)
  - kernel  $k(x, x') = \sum_j \varphi_j^2 \cos [2j\pi(x - x')]$
  - Optimal prediction function  $\theta_*$  has norm  $\|\theta_*\|^2 = \sum_j |\mathcal{F}(\theta_*)_j|^2 \varphi_j^{-2}$
  - Depending on smoothness, may or may not be finite

# From least-squares to non-parametric estimation - II

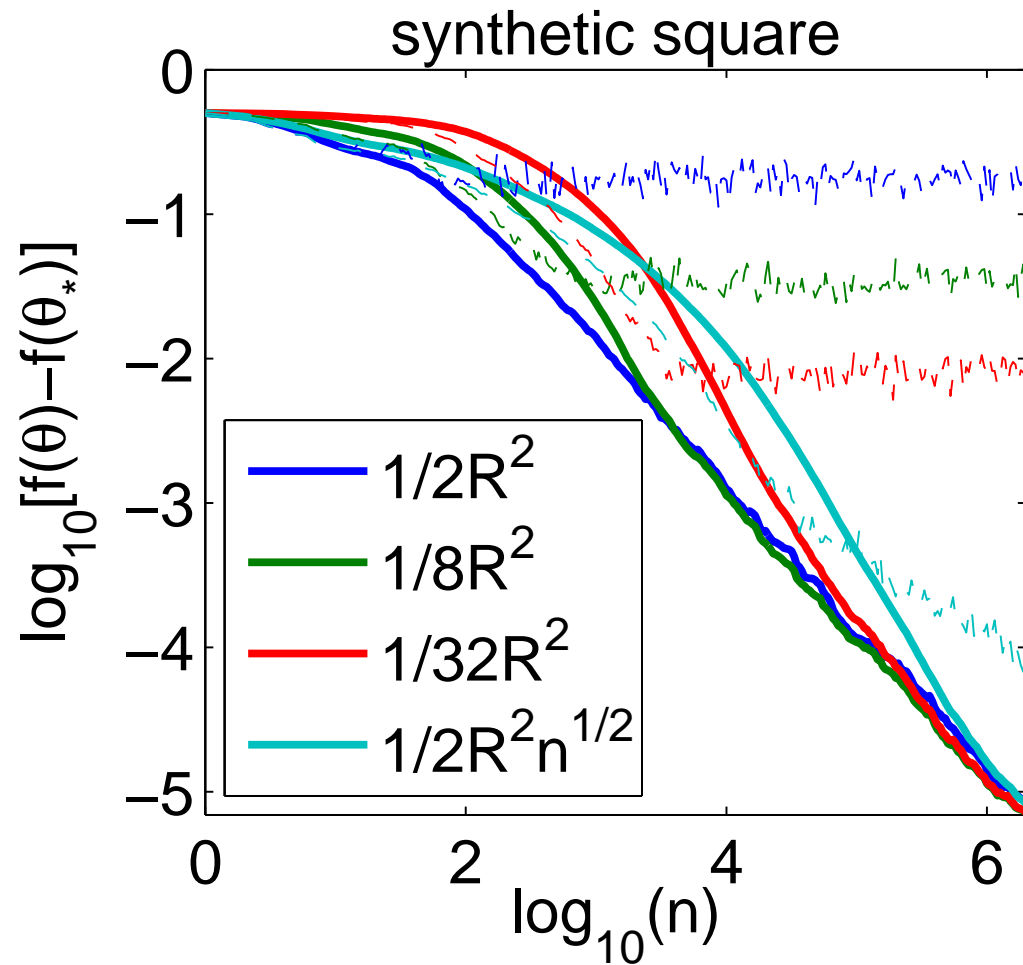
- **Simple example:** Sobolev space on  $\mathcal{X} = [0, 1]$ 
  - $\Phi(x) =$  weighted Fourier basis  $\Phi(x)_j = \varphi_j \cos(2j\pi x)$  (plus sine)
  - kernel  $k(x, x') = \sum_j \varphi_j^2 \cos [2j\pi(x - x')]$
  - Optimal prediction function  $\theta_*$  has norm  $\|\theta_*\|^2 = \sum_j |\mathcal{F}(\theta_*)_j|^2 \varphi_j^{-2}$
  - Depending on smoothness, may or may not be finite
- Adapted norm  $\|H^{1/2-r}\theta_*\|^2 = \sum_j |\mathcal{F}(\theta_*)_j|^2 \varphi_j^{-4r}$  may be finite

$$f(\bar{\theta}_n) - f(\theta_*) \leq \frac{16\sigma^2 \operatorname{tr} H^{1/\alpha}}{n} (\gamma n)^{1/\alpha} + \frac{4 \|H^{1/2-r}(\theta_0 - \theta_*)\|_2}{\gamma^{2r} n^{2 \min\{r, 1\}}}$$

- Same effect than  $\ell_2$ -regularization with weight  $\lambda$  equal to  $\frac{1}{\gamma n}$

# Simulations - synthetic examples

- Gaussian distributions -  $d = 20$



- **Explaining actual behavior for all  $n$**

# Bias-variance decomposition (Défossez and Bach, 2015)

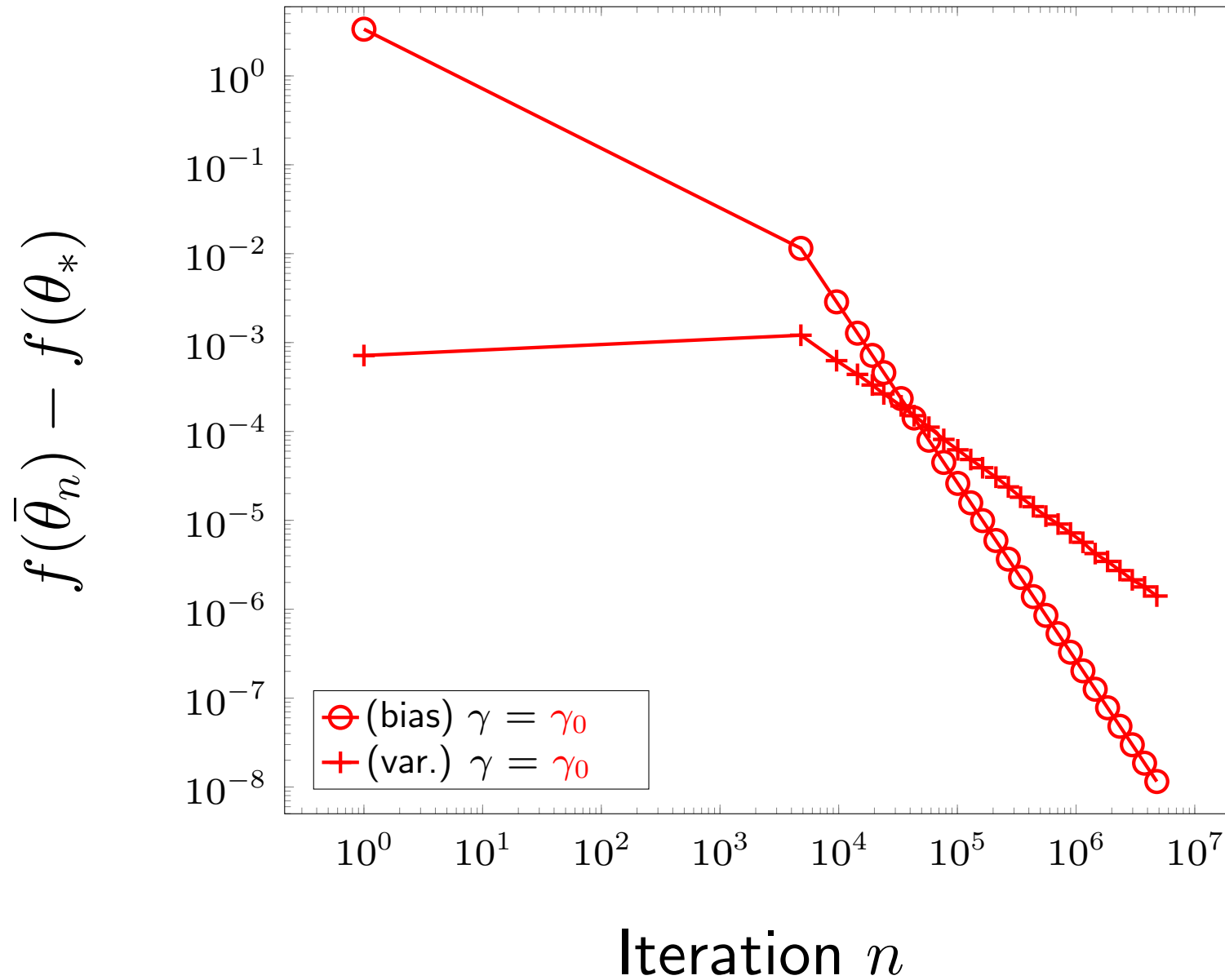
- Simplification: dominating (but exact) term when  $n \rightarrow \infty$  and  $\gamma \rightarrow 0$
- **Variance** (e.g., starting from the solution)

$$f(\bar{\theta}_n) - f(\theta_*) \sim \frac{1}{n} \mathbb{E} \left[ \varepsilon^2 \Phi(x)^\top H^{-1} \Phi(x) \right]$$

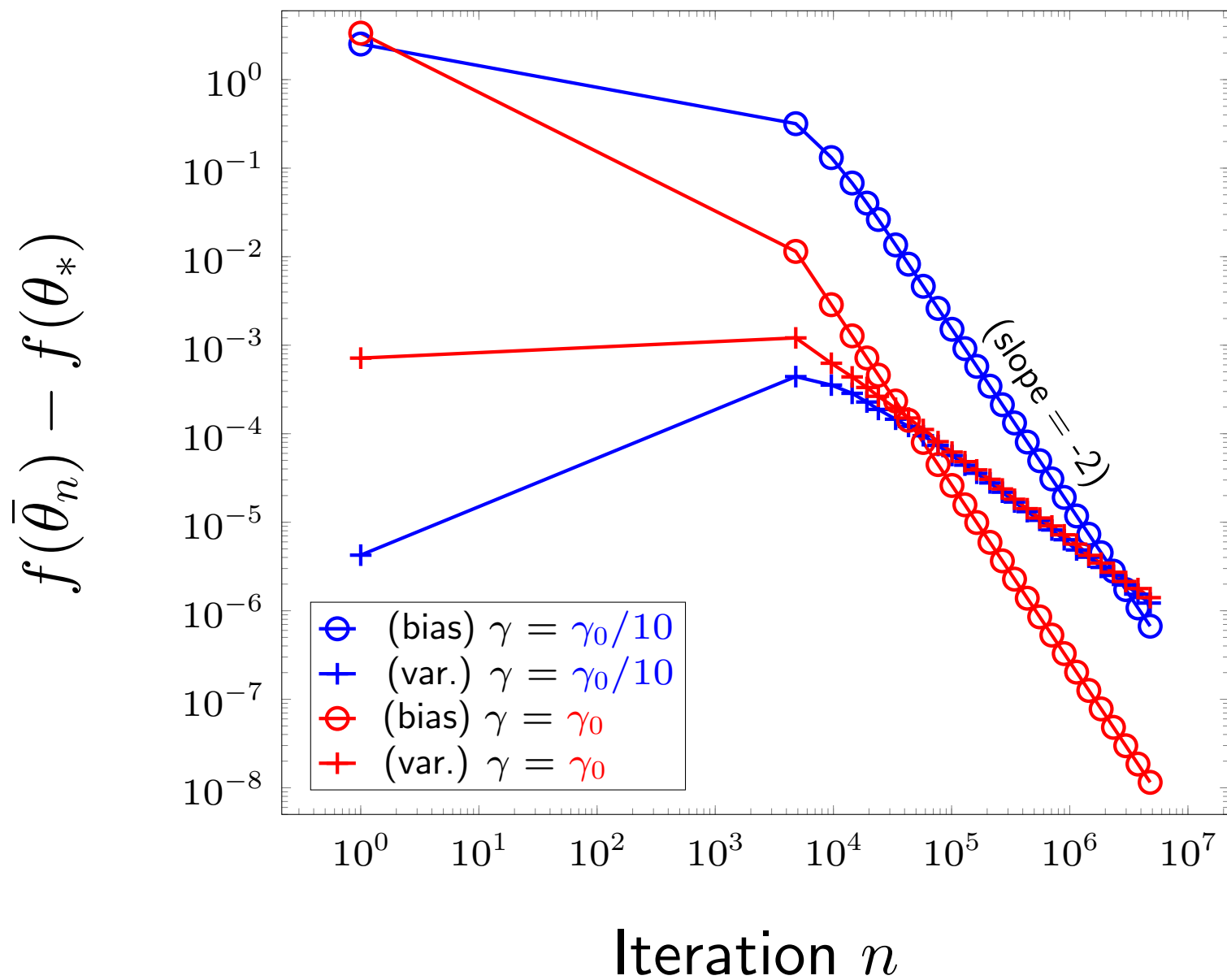
- NB: if noise  $\varepsilon$  is independent, then we obtain  $\frac{d\sigma^2}{n}$
- Exponentially decaying remainder terms (strongly convex problems)
- **Bias** (e.g., no noise)

$$f(\bar{\theta}_n) - f(\theta_*) \sim \frac{1}{n^2 \gamma^2} (\theta_0 - \theta_*)^\top H^{-1} (\theta_0 - \theta_*)$$

# Bias-variance decomposition (synthetic data $d = 25$ )



# Bias-variance decomposition (synthetic data $d = 25$ )



# Optimal sampling (Défossez and Bach, 2015)

- Sampling from a different distribution with importance weights

$$\mathbb{E}_{p(x)p(y|x)} |y - \Phi(x)^\top \theta|^2 = \mathbb{E}_{q(x)p(y|x)} \frac{dp(x)}{dq(x)} |y - \Phi(x)^\top \theta|^2$$

– Recursion:  $\theta_n = \theta_{n-1} - \gamma \frac{dp(x_n)}{dq(x_n)} (\Phi(x_n)^\top \theta_{n-1} - y_n) \Phi(x_n)$

# Optimal sampling (Défossez and Bach, 2015)

- Sampling from a different distribution with importance weights

$$\mathbb{E}_{p(x)p(y|x)} |y - \Phi(x)^\top \theta|^2 = \mathbb{E}_{q(x)p(y|x)} \frac{dp(x)}{dq(x)} |y - \Phi(x)^\top \theta|^2$$

- Recursion:  $\theta_n = \theta_{n-1} - \gamma \frac{dp(x_n)}{dq(x_n)} (\Phi(x_n)^\top \theta_{n-1} - y_n) \Phi(x_n)$
- Specific to least-squares =  $\mathbb{E}_{q(x)p(y|x)} \left| \sqrt{\frac{dp(x)}{dq(x)}} y - \sqrt{\frac{dp(x)}{dq(x)}} \Phi(x)^\top \theta \right|^2$
- Reweighting of the data: same bounds apply!



# Optimal sampling (Défossez and Bach, 2015)

- Sampling from a different distribution with importance weights

$$\mathbb{E}_{p(x)p(y|x)} |y - \Phi(x)^\top \theta|^2 = \mathbb{E}_{q(x)p(y|x)} \frac{dp(x)}{dq(x)} |y - \Phi(x)^\top \theta|^2$$

- Recursion:  $\theta_n = \theta_{n-1} - \gamma \frac{dp(x_n)}{dq(x_n)} (\Phi(x_n)^\top \theta_{n-1} - y_n) \Phi(x_n)$
- Specific to least-squares =  $\mathbb{E}_{q(x)p(y|x)} \left| \sqrt{\frac{dp(x)}{dq(x)}} y - \sqrt{\frac{dp(x)}{dq(x)}} \Phi(x)^\top \theta \right|^2$
- Reweighting of the data: same bounds apply!

- **Optimal for variance:**  $\frac{dq(x)}{dp(x)} \propto \sqrt{\Phi(x)^\top H^{-1} \Phi(x)}$

- Same density as active learning (Kanamori and Shimodaira, 2003)
- Limited gains: different between first and second moments
- Caveat: need to know  $H$

# Optimal sampling (Défossez and Bach, 2015)

- Sampling from a different distribution with importance weights

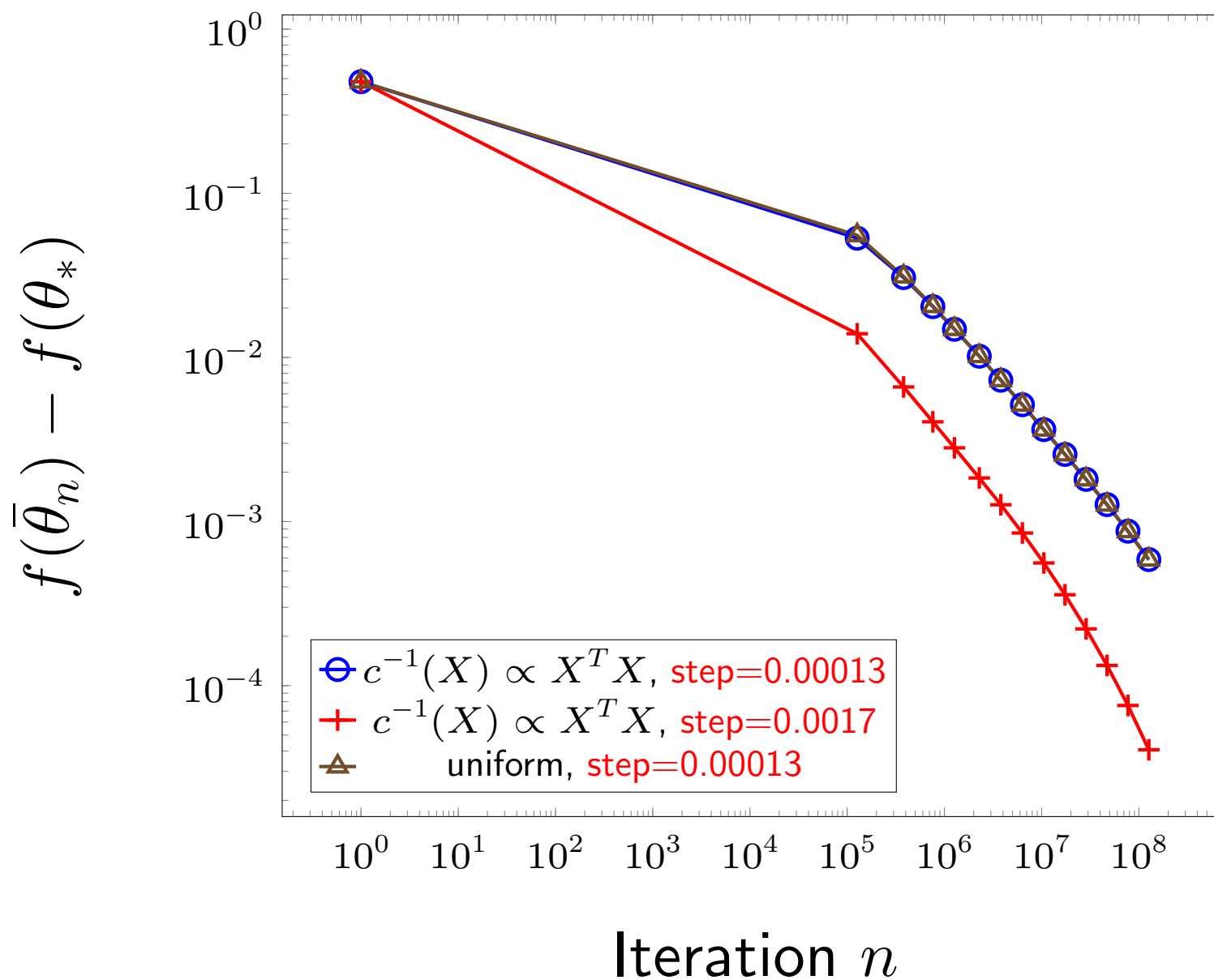
$$\mathbb{E}_{p(x)p(y|x)} |y - \Phi(x)^\top \theta|^2 = \mathbb{E}_{q(x)p(y|x)} \frac{dp(x)}{dq(x)} |y - \Phi(x)^\top \theta|^2$$

- Recursion:  $\theta_n = \theta_{n-1} - \gamma \frac{dp(x_n)}{dq(x_n)} (\Phi(x_n)^\top \theta_{n-1} - y_n) \Phi(x_n)$
- Specific to least-squares =  $\mathbb{E}_{q(x)p(y|x)} \left| \sqrt{\frac{dp(x)}{dq(x)}} y - \sqrt{\frac{dp(x)}{dq(x)}} \Phi(x)^\top \theta \right|^2$
- Reweighting of the data: same bounds apply!

- **Optimal for bias:**  $\frac{dq(x)}{dp(x)} \propto \|\Phi(x)\|^2$

- Simply allows biggest possible step size  $\gamma < \frac{2}{\text{tr } H}$
- Large gains in practice
- Corresponds to normalized least-mean-squares

# Convergence on *Sido* dataset ( $d = 4932$ )



# Achieving optimal bias and variance terms

- Current results with averaged SGD

- **Variance** (starting from optimal  $\theta_*$ ) =  $\frac{\sigma^2 d}{n}$

- **Bias** (no noise) =  $\min \left\{ \frac{R^2 \|\theta_0 - \theta_*\|^2}{n}, \frac{R^4 \langle \theta_0 - \theta_*, H^{-1}(\theta_0 - \theta_*) \rangle}{n^2} \right\}$

# Achieving optimal bias and variance terms

- **Current results with averaged SGD** (ill-conditioned problems)

- **Variance** (starting from optimal  $\theta_*$ ) =  $\frac{\sigma^2 d}{n}$

- **Bias** (no noise) =  $\frac{R^2 \|\theta_0 - \theta_*\|^2}{n}$

# Achieving optimal bias and variance terms

- **Current results with averaged SGD** (ill-conditioned problems)

- **Variance** (starting from optimal  $\theta_*$ ) =  $\frac{\sigma^2 d}{n}$

- **Bias** (no noise) =  $\frac{R^2 \|\theta_0 - \theta_*\|^2}{n}$

	Bias	Variance
<b>Averaged gradient descent</b> (Bach and Moulines, 2013)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n}$	$\frac{\sigma^2 d}{n}$

# Achieving optimal bias and variance terms

	Bias	Variance
<b>Averaged gradient descent</b> (Bach and Moulines, 2013)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n}$	$\frac{\sigma^2 d}{n}$

# Achieving optimal bias and variance terms

	Bias	Variance
<b>Averaged gradient descent</b> (Bach and Moulines, 2013)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n}$	$\frac{\sigma^2 d}{n}$
<b>Accelerated gradient descent</b> (Nesterov, 1983)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n^2}$	$\sigma^2 d$

- **Acceleration is notoriously non-robust to noise** (d'Aspremont, 2008; Schmidt et al., 2011)
  - For non-structured noise, see Lan (2012)



# Achieving optimal bias and variance terms

	Bias	Variance
<b>Averaged gradient descent</b> (Bach and Moulines, 2013)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n}$	$\frac{\sigma^2 d}{n}$
<b>Accelerated gradient descent</b> (Nesterov, 1983)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n^2}$	$\sigma^2 d$
<b>“Between” averaging and acceleration</b> (Flammarion and Bach, 2015)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n^{1+\alpha}}$	$\frac{\sigma^2 d}{n^{1-\alpha}}$

# Achieving optimal bias and variance terms

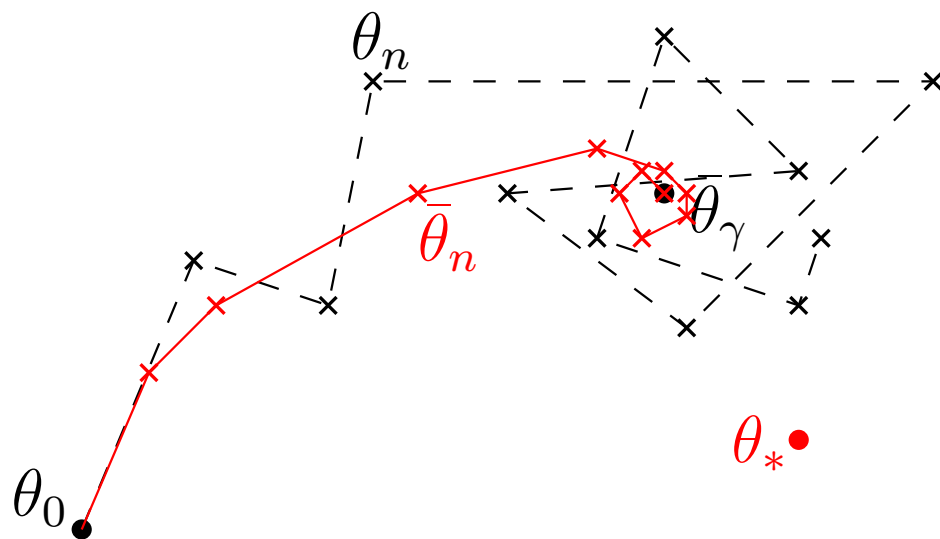
	Bias	Variance
<b>Averaged gradient descent</b> (Bach and Moulines, 2013)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n}$	$\frac{\sigma^2 d}{n}$
<b>Accelerated gradient descent</b> (Nesterov, 1983)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n^2}$	$\sigma^2 d$
<b>“Between” averaging and acceleration</b> (Flammarion and Bach, 2015)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n^{1+\alpha}}$	$\frac{\sigma^2 d}{n^{1-\alpha}}$
<b>Averaging and acceleration</b> (Dieuleveut, Flammarion, and Bach, 2016)	$\frac{R^2 \ \theta_0 - \theta_*\ ^2}{n^2}$	$\frac{\sigma^2 d}{n}$

# Beyond least-squares - Markov chain interpretation

- Recursion  $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$  also defines a Markov chain
  - Stationary distribution  $\pi_\gamma$  such that  $\int f'(\theta)\pi_\gamma(d\theta) = 0$
  - When  $f'$  is not linear,  $f'(\int \theta\pi_\gamma(d\theta)) \neq \int f'(\theta)\pi_\gamma(d\theta) = 0$

# Beyond least-squares - Markov chain interpretation

- Recursion  $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$  also defines a Markov chain
  - Stationary distribution  $\pi_\gamma$  such that  $\int f'(\theta)\pi_\gamma(d\theta) = 0$
  - When  $f'$  is not linear,  $f'(\int \theta\pi_\gamma(d\theta)) \neq \int f'(\theta)\pi_\gamma(d\theta) = 0$
- $\theta_n$  oscillates around the wrong value  $\bar{\theta}_\gamma \neq \theta_*$

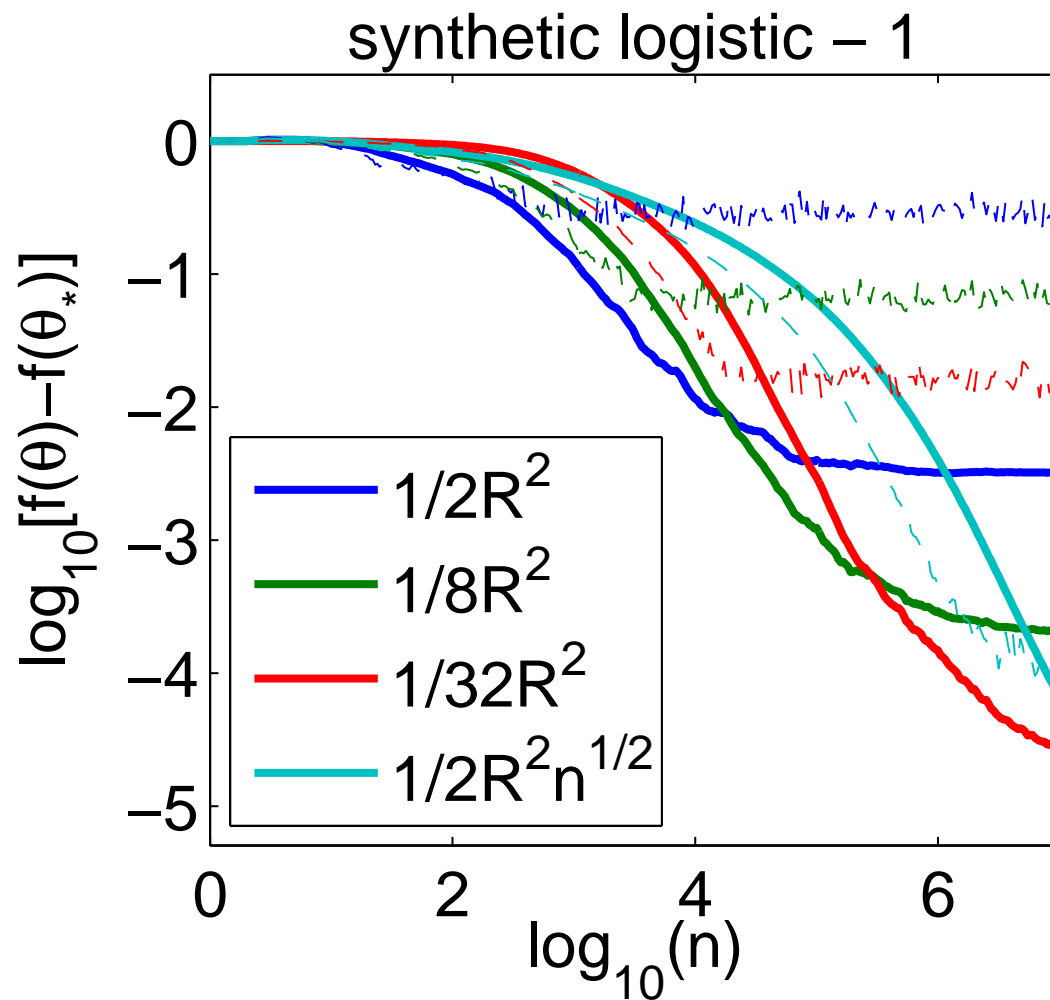


# Beyond least-squares - Markov chain interpretation

- Recursion  $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$  also defines a Markov chain
  - Stationary distribution  $\pi_\gamma$  such that  $\int f'(\theta)\pi_\gamma(d\theta) = 0$
  - When  $f'$  is not linear,  $f'(\int \theta\pi_\gamma(d\theta)) \neq \int f'(\theta)\pi_\gamma(d\theta) = 0$
- $\theta_n$  oscillates around the wrong value  $\bar{\theta}_\gamma \neq \theta_*$ 
  - moreover,  $\|\theta_* - \theta_n\| = O_p(\sqrt{\gamma})$
  - Linear convergence up to the noise level for strongly-convex problems (Nedic and Bertsekas, 2000)
- Ergodic theorem
  - averaged iterates converge to  $\bar{\theta}_\gamma \neq \theta_*$  at rate  $O(1/n)$
  - moreover,  $\|\theta_* - \bar{\theta}_\gamma\| = O(\gamma)$  (Bach, 2013)

# Simulations - synthetic examples

- Gaussian distributions -  $d = 20$



# Restoring convergence through online Newton steps

- **Known facts**

1. Averaged SGD with  $\gamma_n \propto n^{-1/2}$  leads to *robust* rate  $O(n^{-1/2})$  for all convex functions
2. Averaged SGD with  $\gamma_n$  constant leads to *robust* rate  $O(n^{-1})$  for all convex *quadratic* functions
3. Newton's method squares the error at each iteration for smooth functions
4. A single step of Newton's method is equivalent to minimizing the quadratic Taylor expansion

# Restoring convergence through online Newton steps

- **Known facts**

1. Averaged SGD with  $\gamma_n \propto n^{-1/2}$  leads to *robust* rate  $O(n^{-1/2})$  for all convex functions
2. Averaged SGD with  $\gamma_n$  constant leads to *robust* rate  $O(n^{-1})$  for all convex *quadratic* functions  $\Rightarrow O(n^{-1})$
3. Newton's method squares the error at each iteration for smooth functions  $\Rightarrow O((n^{-1/2})^2)$
4. A single step of Newton's method is equivalent to minimizing the quadratic Taylor expansion

- **Online Newton step**

- Rate:  $O((n^{-1/2})^2 + n^{-1}) = O(n^{-1})$
- Complexity:  $O(d)$  per iteration



# Restoring convergence through online Newton steps

- The Newton step for  $f = \mathbb{E}f_n(\theta) \stackrel{\text{def}}{=} \mathbb{E}[\ell(y_n, \langle \theta, \Phi(x_n) \rangle)]$  at  $\tilde{\theta}$  is equivalent to minimizing the quadratic approximation

$$\begin{aligned} g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= f(\tilde{\theta}) + \langle \mathbb{E}f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, \mathbb{E}f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= \mathbb{E} \left[ f(\tilde{\theta}) + \langle f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \right] \end{aligned}$$

# Restoring convergence through online Newton steps

- The Newton step for  $f = \mathbb{E}f_n(\theta) \stackrel{\text{def}}{=} \mathbb{E}[\ell(y_n, \langle \theta, \Phi(x_n) \rangle)]$  at  $\tilde{\theta}$  is equivalent to minimizing the quadratic approximation

$$\begin{aligned}g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= f(\tilde{\theta}) + \langle \mathbb{E}f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, \mathbb{E}f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \\ &= \mathbb{E} \left[ f(\tilde{\theta}) + \langle f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \right]\end{aligned}$$

- **Complexity of least-mean-square recursion for  $g$  is  $O(d)$**

$$\theta_n = \theta_{n-1} - \gamma [f'_n(\tilde{\theta}) + f''_n(\tilde{\theta})(\theta_{n-1} - \tilde{\theta})]$$

- $f''_n(\tilde{\theta}) = \ell''(y_n, \langle \tilde{\theta}, \Phi(x_n) \rangle) \Phi(x_n) \otimes \Phi(x_n)$  has rank one
- **New online Newton step without computing/inverting Hessians**

# Choice of support point for online Newton step

- **Two-stage procedure**

- (1) Run  $n/2$  iterations of averaged SGD to obtain  $\tilde{\theta}$
- (2) Run  $n/2$  iterations of averaged constant step-size LMS
  - Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)
  - **Provable convergence rate of  $O(d/n)$**  for logistic regression
  - Additional assumptions but no **strong convexity**

# Logistic regression - Proof technique

- Using generalized self-concordance of  $\varphi : u \mapsto \log(1 + e^{-u})$ :

$$|\varphi'''(u)| \leq \varphi''(u)$$

- NB: difference with regular self-concordance:  $|\varphi'''(u)| \leq 2\varphi''(u)^{3/2}$
- Using novel high-probability convergence results for regular averaged stochastic gradient descent
- Requires assumption on the kurtosis in every direction, i.e.,

$$\mathbb{E}\langle \Phi(x_n), \eta \rangle^4 \leq \kappa [\mathbb{E}\langle \Phi(x_n), \eta \rangle^2]^2$$

# Choice of support point for online Newton step

- **Two-stage procedure**

- (1) Run  $n/2$  iterations of averaged SGD to obtain  $\tilde{\theta}$

- (2) Run  $n/2$  iterations of averaged constant step-size LMS

- Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)

- **Provable convergence rate of  $O(d/n)$**  for logistic regression

- Additional assumptions but no **strong convexity**

- **Update at each iteration using the current averaged iterate**

- Recursion: 
$$\theta_n = \theta_{n-1} - \gamma [f'_n(\bar{\theta}_{n-1}) + f''_n(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1})]$$

- No provable convergence rate (yet) but best practical behavior

- Note (dis)similarity with regular SGD:  $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$

# Online Newton algorithm

## Current proof (Flammarion et al., 2014)

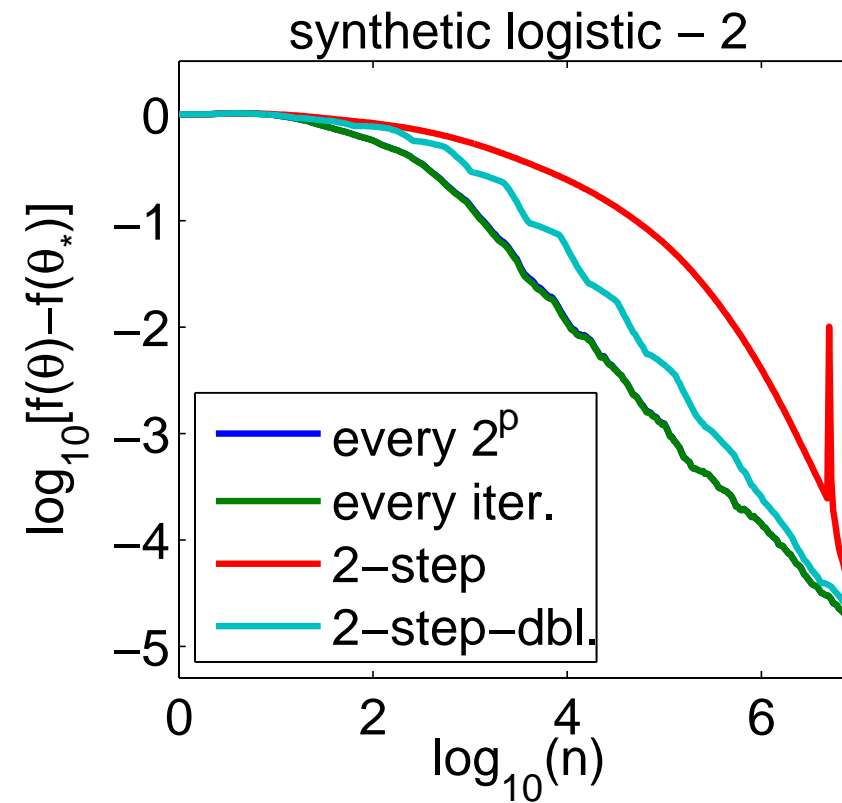
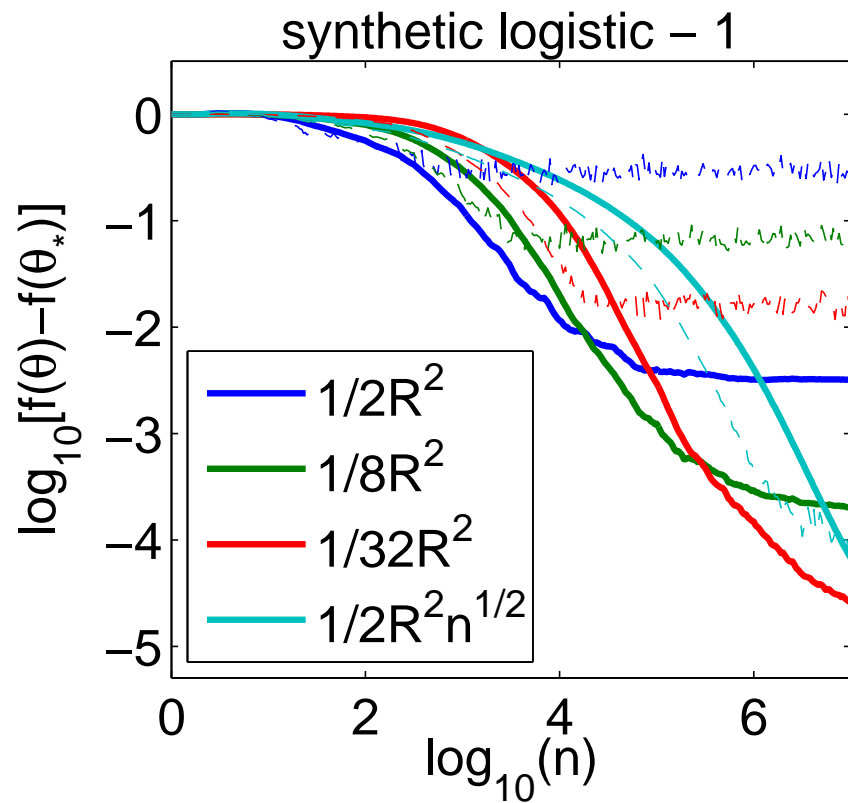
- Recursion

$$\begin{cases} \theta_n &= \theta_{n-1} - \gamma [f'_n(\bar{\theta}_{n-1}) + f''_n(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1})] \\ \bar{\theta}_n &= \bar{\theta}_{n-1} + \frac{1}{n}(\theta_n - \bar{\theta}_{n-1}) \end{cases}$$

- Instance of **two-time-scale** stochastic approximation (Borkar, 1997)
  - Given  $\bar{\theta}$ ,  $\theta_n = \theta_{n-1} - \gamma [f'_n(\bar{\theta}) + f''_n(\bar{\theta})(\theta_{n-1} - \bar{\theta})]$  defines a homogeneous Markov chain (fast dynamics)
  - $\bar{\theta}_n$  is updated at rate  $1/n$  (slow dynamics)
- **Difficulty:** preserving robustness to ill-conditioning

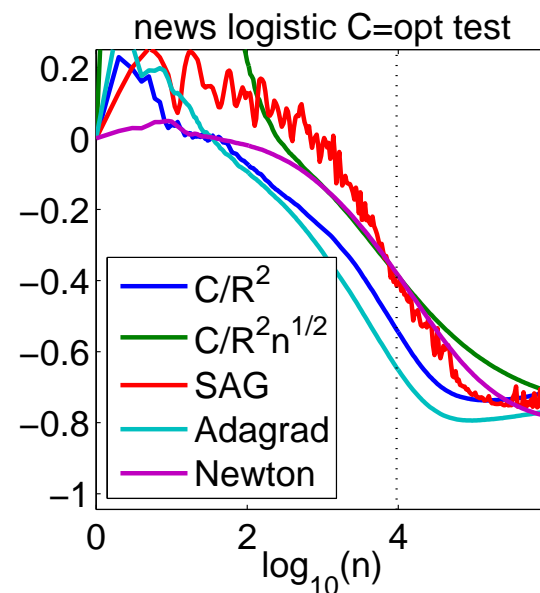
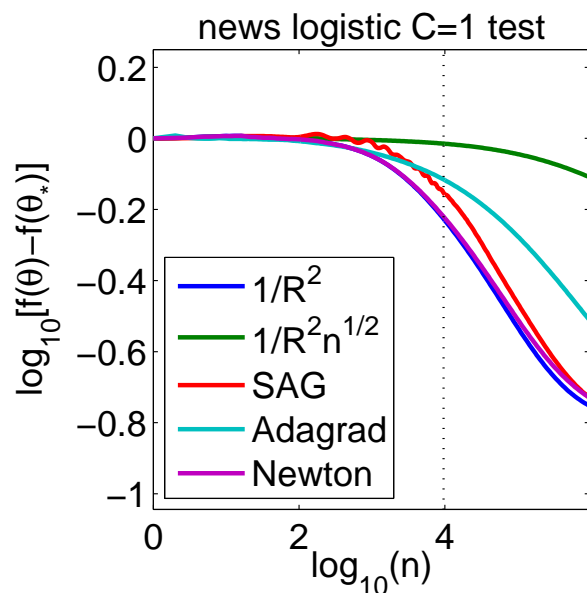
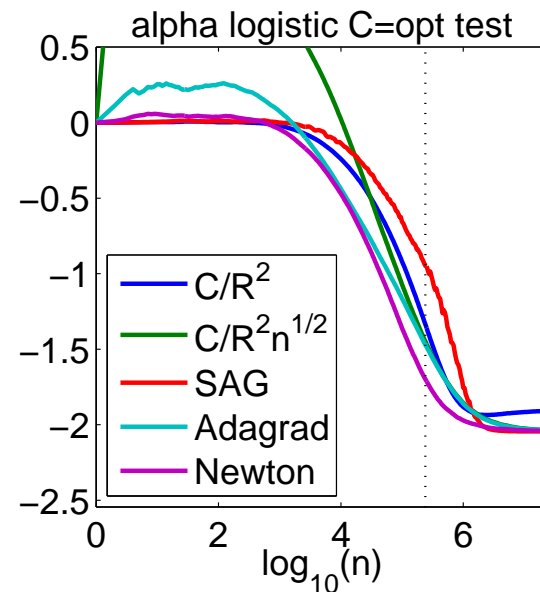
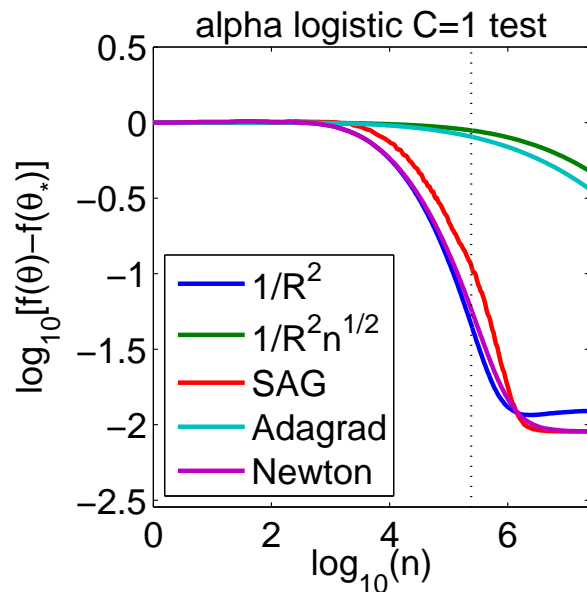
# Simulations - synthetic examples

- Gaussian distributions -  $d = 20$



# Simulations - benchmarks

- *alpha* ( $d = 500, n = 500\,000$ ), *news* ( $d = 1\,300\,000, n = 20\,000$ )





# Why is $\frac{\sigma^2 d}{n}$ optimal for least-squares?

- Reduction to an hypothesis testing problem
  - Application of Varshamov-Gilbert's lemma
- **Best possible prediction independently of computation**
  - To be contrasted with lower bounds based on specific models of computation
- See <http://www-math.mit.edu/~rigollet/PDFs/RigNotes15.pdf>

# Summary of rates of convergence

- Problem parameters
  - $D$  diameter of the domain
  - $B$  Lipschitz-constant
  - $L$  smoothness constant
  - $\mu$  strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$ stochastic: $BD/\sqrt{n}$	deterministic: $B^2/(t\mu)$ stochastic: $B^2/(n\mu)$
smooth	deterministic: $LD^2/t^2$ stochastic: $LD^2/\sqrt{n}$	deterministic: $\exp(-t\sqrt{\mu/L})$ stochastic: $L/(n\mu)$
quadratic	deterministic: $LD^2/t^2$ stochastic: $d/n + LD^2/n$	deterministic: $\exp(-t\sqrt{\mu/L})$ stochastic: $d/n + LD^2/n$

# Summary of rates of convergence

- Problem parameters
  - $D$  diameter of the domain
  - $B$  Lipschitz-constant
  - $L$  smoothness constant
  - $\mu$  strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$ stochastic: $BD/\sqrt{n}$	deterministic: $B^2/(t\mu)$ stochastic: $B^2/(n\mu)$
smooth	deterministic: $LD^2/t^2$ stochastic: $LD^2/\sqrt{n}$ finite sum: $n/t$	deterministic: $\exp(-t\sqrt{\mu/L})$ stochastic: $L/(n\mu)$ finite sum: $\exp(-\min\{1/n, \mu/L\}t)$
quadratic	deterministic: $LD^2/t^2$ stochastic: $d/n + LD^2/n$	deterministic: $\exp(-t\sqrt{\mu/L})$ stochastic: $d/n + LD^2/n$

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex

# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# Going beyond a single pass over the data

- **Stochastic approximation**

- Assumes infinite data stream
- Observations are used only once
- Directly minimizes **testing** cost  $\mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$

# Going beyond a single pass over the data

- **Stochastic approximation**

- Assumes infinite data stream
- Observations are used only once
- Directly minimizes **testing** cost  $\mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$

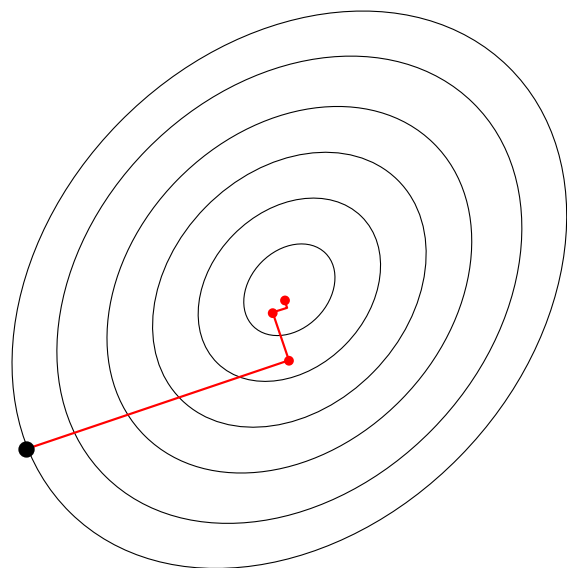
- **Machine learning practice**

- Finite data set  $(x_1, y_1, \dots, x_n, y_n)$
- Multiple passes
- Minimizes **training** cost  $\frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
- Need to regularize (e.g., by the  $\ell_2$ -norm) to avoid overfitting

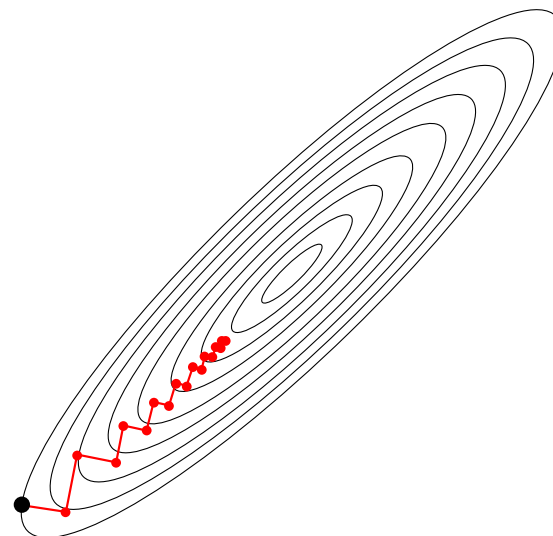
- **Goal:** minimize  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$

# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  **convex** and  $L$ -smooth on  $\mathbb{R}^d$
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$



(small  $\kappa = L/\mu$ )



(large  $\kappa = L/\mu$ )



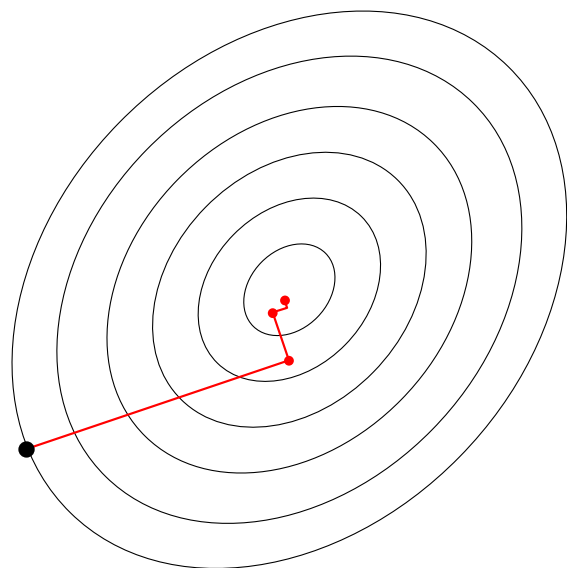
# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  **convex** and  $L$ -smooth on  $\mathbb{R}^d$

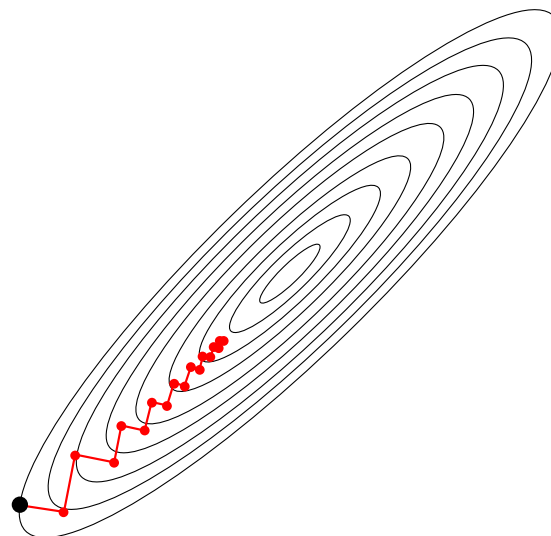
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$

$$g(\theta_t) - g(\theta_*) \leq O(1/t)$$

$$g(\theta_t) - g(\theta_*) \leq O((1 - \mu/L)^t) = O(e^{-t(\mu/L)}) \text{ if } \mu\text{-strongly convex}$$



(small  $\kappa = L/\mu$ )



(large  $\kappa = L/\mu$ )

# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  convex and  $L$ -smooth on  $\mathbb{R}^d$
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for convex functions
  - $O(e^{-t/\kappa})$  *linear* if strongly-convex
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1}g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  *quadratic* rate

# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  convex and  $L$ -smooth on  $\mathbb{R}^d$
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for convex functions
  - $O(e^{-t/\kappa})$  *linear* if strongly-convex  $\Leftrightarrow O(\kappa \log \frac{1}{\varepsilon})$  iterations
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  *quadratic* rate  $\Leftrightarrow O(\log \log \frac{1}{\varepsilon})$  iterations

# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  convex and  $L$ -smooth on  $\mathbb{R}^d$
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for convex functions
  - $O(e^{-t/\kappa})$  *linear* if strongly-convex  $\Leftrightarrow$  **complexity** =  $O(nd \cdot \kappa \log \frac{1}{\epsilon})$
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  *quadratic* rate  $\Leftrightarrow$  **complexity** =  $O((nd^2 + d^3) \cdot \log \log \frac{1}{\epsilon})$

# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  convex and  $L$ -smooth on  $\mathbb{R}^d$
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for convex functions
  - $O(e^{-t/\kappa})$  *linear* if strongly-convex  $\Leftrightarrow$  **complexity** =  $O(nd \cdot \kappa \log \frac{1}{\epsilon})$
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  *quadratic* rate  $\Leftrightarrow$  **complexity** =  $O((nd^2 + d^3) \cdot \log \log \frac{1}{\epsilon})$
- **Key insights for machine learning (Bottou and Bousquet, 2008)**
  1. No need to optimize below statistical error
  2. Cost functions are averages
  3. Testing error is more important than training error

# Iterative methods for minimizing smooth functions

- **Assumption:**  $g$  convex and  $L$ -smooth on  $\mathbb{R}^d$
- **Gradient descent:**  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$ 
  - $O(1/t)$  convergence rate for convex functions
  - $O(e^{-t/\kappa})$  linear if strongly-convex  $\Leftrightarrow$  complexity =  $O(nd \cdot \kappa \log \frac{1}{\epsilon})$
- **Newton method:**  $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$ 
  - $O(e^{-\rho 2^t})$  quadratic rate  $\Leftrightarrow$  complexity =  $O((nd^2 + d^3) \cdot \log \log \frac{1}{\epsilon})$
- **Key insights for machine learning (Bottou and Bousquet, 2008)**
  1. No need to optimize below statistical error
  2. Cost functions are averages
  3. Testing error is more important than training error

# Stochastic gradient descent (SGD) for finite sums

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$$

- **Iteration:**  $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$ 
  - Sampling with replacement:  $i(t)$  random element of  $\{1, \dots, n\}$
  - Polyak-Ruppert averaging:  $\bar{\theta}_t = \frac{1}{t+1} \sum_{u=0}^t \theta_u$

# Stochastic gradient descent (SGD) for finite sums

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$$

- **Iteration:**  $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$ 
  - Sampling with replacement:  $i(t)$  random element of  $\{1, \dots, n\}$
  - Polyak-Ruppert averaging:  $\bar{\theta}_t = \frac{1}{t+1} \sum_{u=0}^t \theta_u$
- **Convergence rate** if each  $f_i$  is convex  $L$ -smooth and  $g$   $\mu$ -strongly-convex:

$$\mathbb{E}g(\bar{\theta}_t) - g(\theta_*) \leq \begin{cases} O(1/\sqrt{t}) & \text{if } \gamma_t = 1/(L\sqrt{t}) \\ O(L/(\mu t)) = O(\kappa/t) & \text{if } \gamma_t = 1/(\mu t) \end{cases}$$

- No adaptivity to strong-convexity in general
- Adaptivity with self-concordance assumption (Bach, 2013)
- Running-time complexity:  $O(d \cdot \kappa/\varepsilon)$



## Stochastic vs. deterministic methods

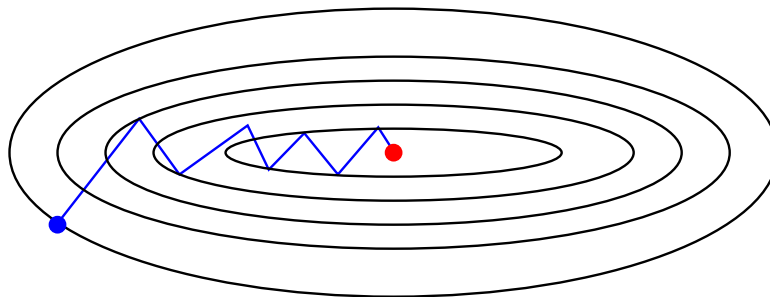
- Minimizing  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$  with  $f_i(\theta) = \ell(y_i, h(x_i, \theta)) + \lambda \Omega(\theta)$

# Stochastic vs. deterministic methods

- Minimizing  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$  with  $f_i(\theta) = \ell(y_i, h(x_i, \theta)) + \lambda \Omega(\theta)$
- **Batch** gradient descent:  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n f'_i(\theta_{t-1})$ 
  - Linear (e.g., exponential) convergence rate in  $O(e^{-t/\kappa})$
  - Iteration complexity is linear in  $n$

# Stochastic vs. deterministic methods

- Minimizing  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$  with  $f_i(\theta) = \ell(y_i, h(x_i, \theta)) + \lambda \Omega(\theta)$
- **Batch** gradient descent:  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n f'_i(\theta_{t-1})$



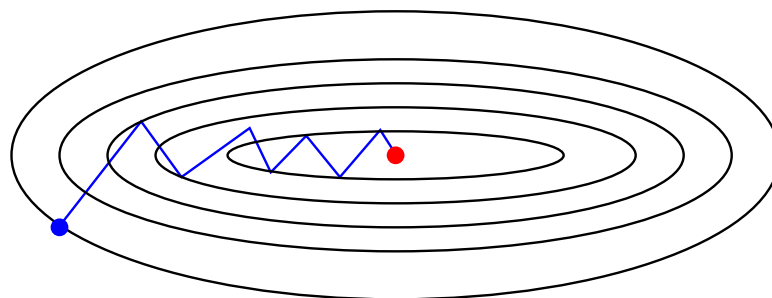
# Stochastic vs. deterministic methods

- Minimizing  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$  with  $f_i(\theta) = \ell(y_i, h(x_i, \theta)) + \lambda \Omega(\theta)$
- **Batch** gradient descent:  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n f'_i(\theta_{t-1})$ 
  - Linear (e.g., exponential) convergence rate in  $O(e^{-t/\kappa})$
  - Iteration complexity is linear in  $n$
- **Stochastic** gradient descent:  $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$ 
  - Sampling with replacement:  $i(t)$  random element of  $\{1, \dots, n\}$
  - Convergence rate in  $O(\kappa/t)$
  - Iteration complexity is independent of  $n$

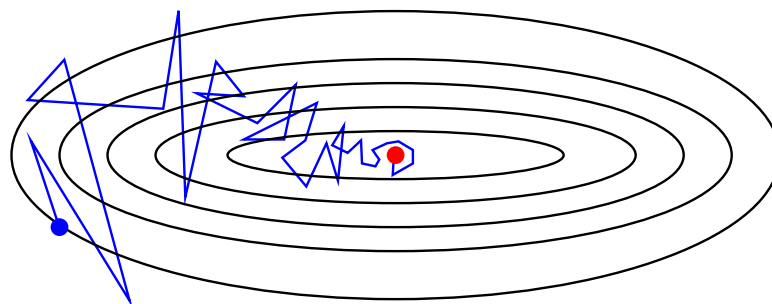
# Stochastic vs. deterministic methods

- Minimizing  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$  with  $f_i(\theta) = \ell(y_i, h(x_i, \theta)) + \lambda\Omega(\theta)$

- **Batch** gradient descent:  $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n f'_i(\theta_{t-1})$

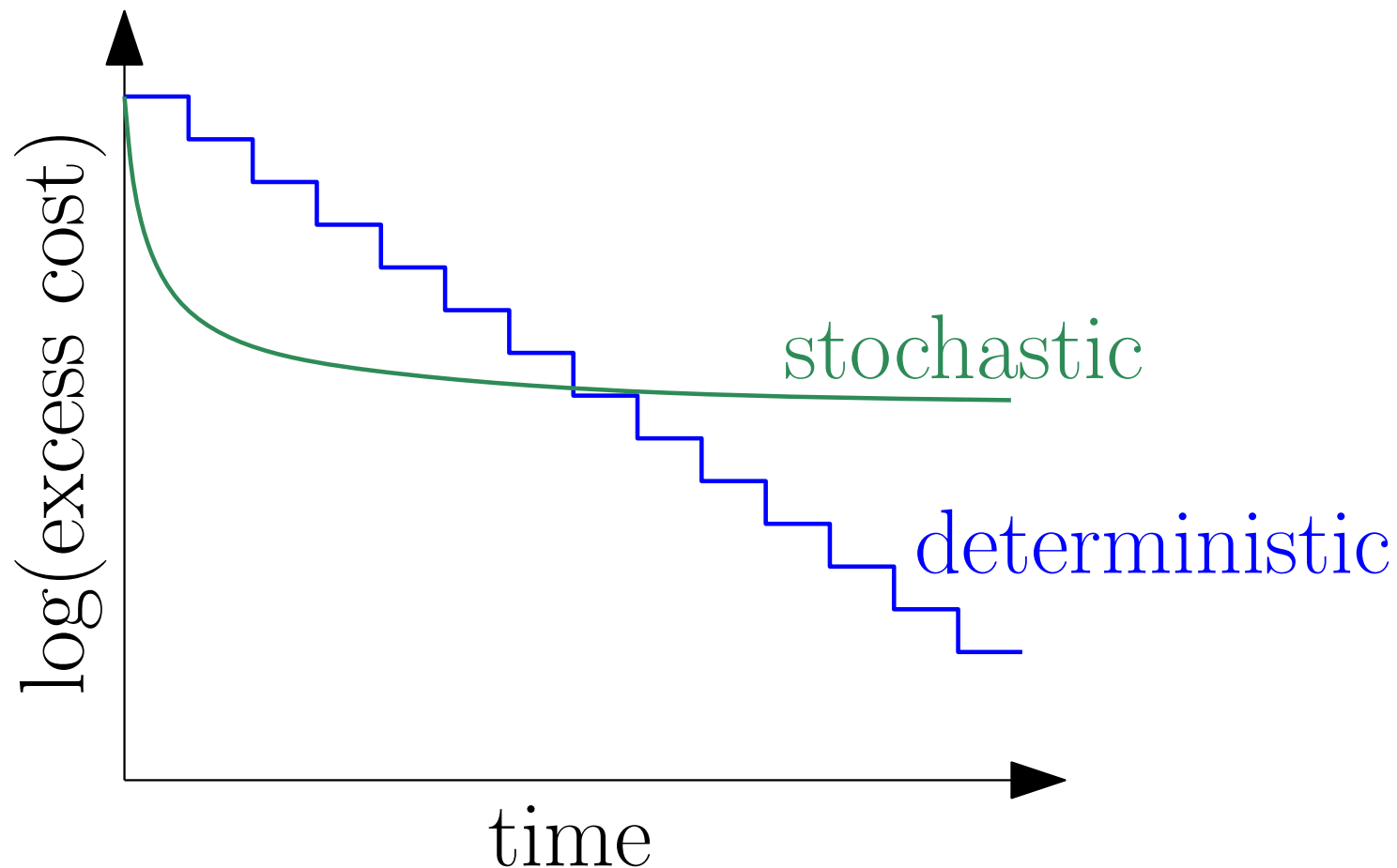


- **Stochastic** gradient descent:  $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$



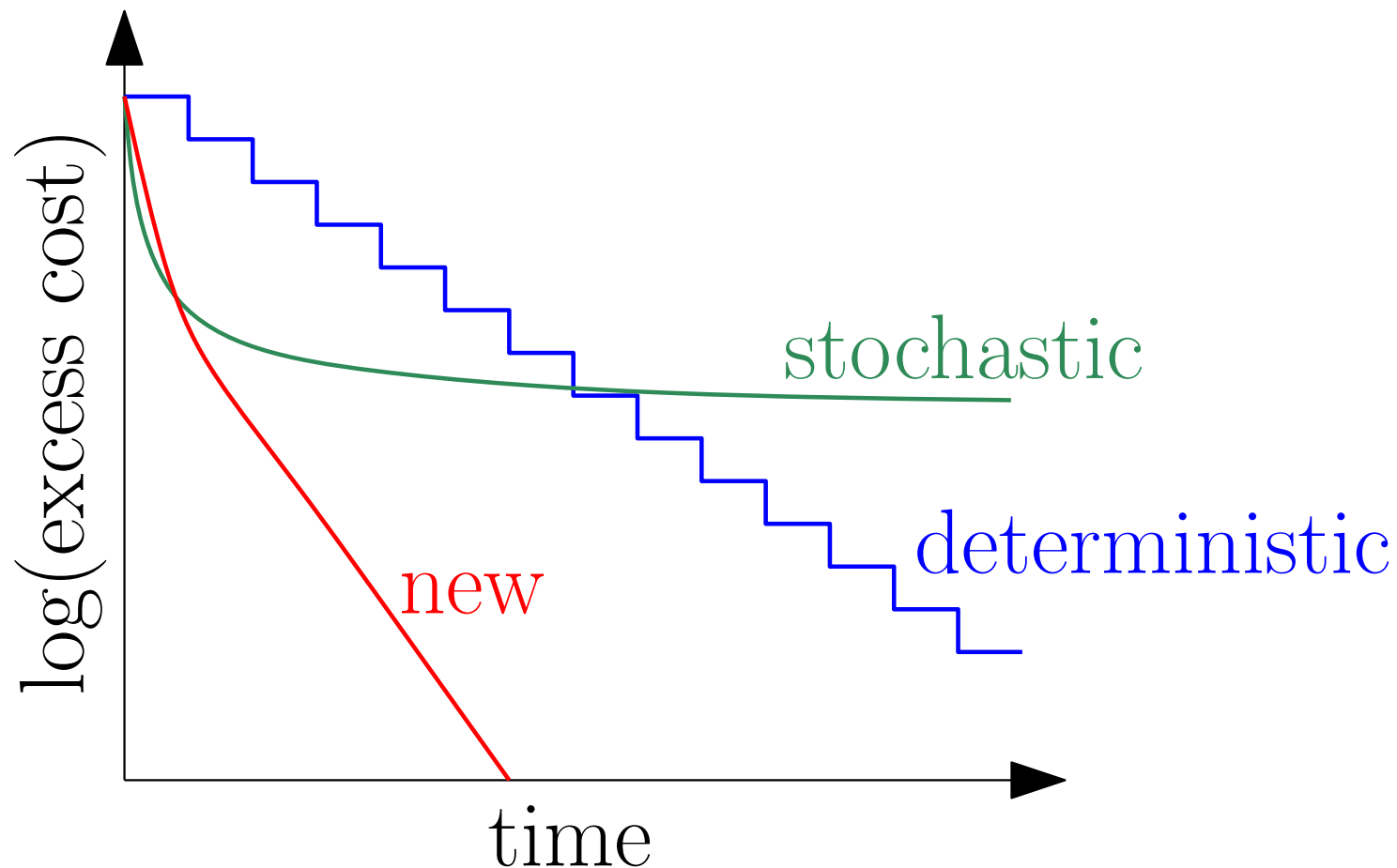
# Stochastic vs. deterministic methods

- **Goal** = best of both worlds: Linear rate with  $O(d)$  iteration cost  
Simple choice of step size



# Stochastic vs. deterministic methods

- **Goal** = best of both worlds: Linear rate with  $O(d)$  iteration cost  
Simple choice of step size



# Accelerating gradient methods - Related work

- **Generic acceleration** (Nesterov, 1983, 2004)

$$\theta_t = \eta_{t-1} - \gamma_t g'(\eta_{t-1}) \text{ and } \eta_t = \theta_t + \delta_t(\theta_t - \theta_{t-1})$$



# Accelerating gradient methods - Related work

- **Generic acceleration** (Nesterov, 1983, 2004)

$$\theta_t = \eta_{t-1} - \gamma_t g'(\eta_{t-1}) \text{ and } \eta_t = \theta_t + \delta_t(\theta_t - \theta_{t-1})$$

- Good choice of momentum term  $\delta_t \in [0, 1)$

$$g(\theta_t) - g(\theta_*) \leq O(1/t^2)$$

$$g(\theta_t) - g(\theta_*) \leq O(e^{-t\sqrt{\mu/L}}) = O(e^{-t/\sqrt{\kappa}}) \text{ if } \mu\text{-strongly convex}$$

- **Optimal rates** after  $t = O(d)$  iterations (Nesterov, 2004)

# Accelerating gradient methods - Related work

- **Generic acceleration** (Nesterov, 1983, 2004)

$$\theta_t = \eta_{t-1} - \gamma_t g'(\eta_{t-1}) \text{ and } \eta_t = \theta_t + \delta_t(\theta_t - \theta_{t-1})$$

- Good choice of momentum term  $\delta_t \in [0, 1)$

$$g(\theta_t) - g(\theta_*) \leq O(1/t^2)$$

$$g(\theta_t) - g(\theta_*) \leq O(e^{-t\sqrt{\mu/L}}) = O(e^{-t/\sqrt{\kappa}}) \text{ if } \mu\text{-strongly convex}$$

- **Optimal rates** after  $t = O(d)$  iterations (Nesterov, 2004)

- Still  $O(nd)$  iteration cost: complexity =  $O(nd \cdot \sqrt{\kappa} \log \frac{1}{\epsilon})$

# Accelerating gradient methods - Related work

- **Constant step-size stochastic gradient**
  - Solodov (1998); Nedic and Bertsekas (2000)
  - Linear convergence, but only up to a fixed tolerance

# Accelerating gradient methods - Related work

- **Constant step-size stochastic gradient**
  - Solodov (1998); Nedic and Bertsekas (2000)
  - Linear convergence, but only up to a fixed tolerance
- **Stochastic methods in the dual (SDCA)**
  - Shalev-Shwartz and Zhang (2012)
  - Similar linear rate but limited choice for the  $f_i$ 's
  - Extensions without duality: see Shalev-Shwartz (2016)

# Accelerating gradient methods - Related work

- **Constant step-size stochastic gradient**
  - Solodov (1998); Nedic and Bertsekas (2000)
  - Linear convergence, but only up to a fixed tolerance
- **Stochastic methods in the dual (SDCA)**
  - Shalev-Shwartz and Zhang (2012)
  - Similar linear rate but limited choice for the  $f_i$ 's
  - Extensions without duality: see Shalev-Shwartz (2016)
- **Stochastic version of accelerated batch gradient methods**
  - Tseng (1998); Ghadimi and Lan (2010); Xiao (2010)
  - Can improve constants, but still have sublinear  $O(1/t)$  rate

# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**
  - Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$
  - Random selection  $i(t) \in \{1, \dots, n\}$  with replacement
  - Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**

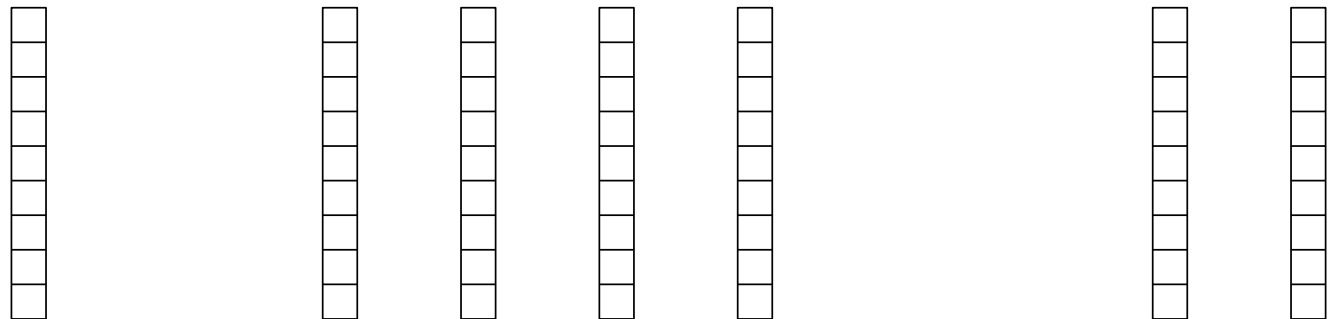
- Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$

- Random selection  $i(t) \in \{1, \dots, n\}$  with replacement

- Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

functions  $g = \frac{1}{n} \sum_{i=1}^n f_i$      $f_1$      $f_2$      $f_3$      $f_4$      $\dots$      $f_{n-1}$      $f_n$

gradients  $\in \mathbb{R}^d$      $\frac{1}{n} \sum_{i=1}^n y_i^t$      $y_1^t$      $y_2^t$      $y_3^t$      $y_4^t$      $\dots$      $y_{n-1}^t$      $y_n^t$



# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**

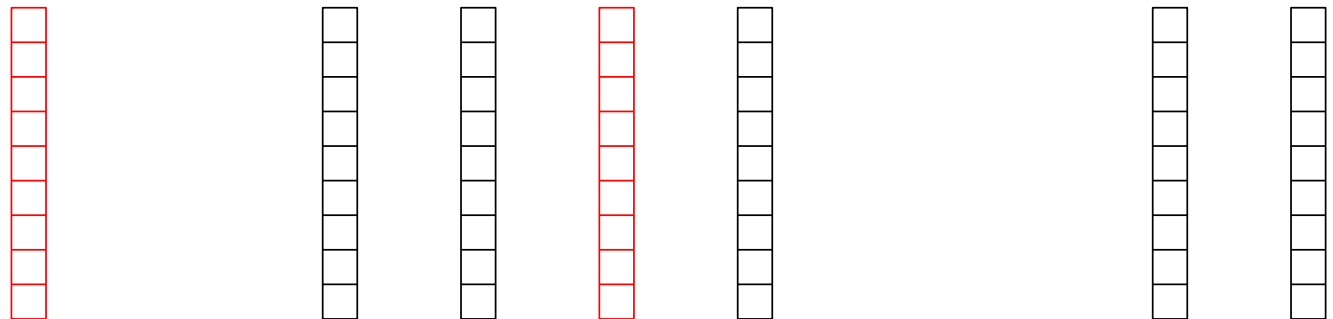
- Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$

- Random selection  $i(t) \in \{1, \dots, n\}$  with replacement

- Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

functions  $g = \frac{1}{n} \sum_{i=1}^n f_i$      $f_1$      $f_2$      $f_3$      $f_4$      $\dots$      $f_{n-1}$      $f_n$

gradients  $\in \mathbb{R}^d$      $\frac{1}{n} \sum_{i=1}^n y_i^t$      $y_1^t$      $y_2^t$      $y_3^t$      $y_4^t$      $\dots$      $y_{n-1}^t$      $y_n^t$





# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**

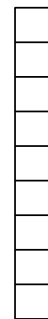
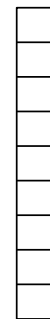
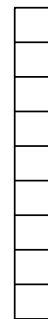
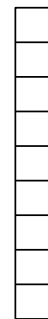
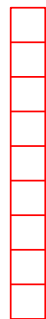
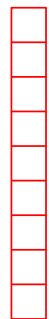
- Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$

- Random selection  $i(t) \in \{1, \dots, n\}$  with replacement

- Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

functions  $g = \frac{1}{n} \sum_{i=1}^n f_i$      $f_1$      $f_2$      $f_3$      $f_4$      $\dots$      $f_{n-1}$      $f_n$

gradients  $\in \mathbb{R}^d$      $\frac{1}{n} \sum_{i=1}^n y_i^t$      $y_1^t$      $y_2^t$      $y_3^t$      $y_4^t$      $\dots$      $y_{n-1}^t$      $y_n^t$



# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**
  - Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$
  - Random selection  $i(t) \in \{1, \dots, n\}$  with replacement
  - Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$
- Stochastic version of incremental average gradient (Blatt et al., 2008)

# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**
  - Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$
  - Random selection  $i(t) \in \{1, \dots, n\}$  with replacement
  - Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$
- Stochastic version of incremental average gradient (Blatt et al., 2008)
- **Extra memory requirement:**  $n$  gradients in  $\mathbb{R}^d$  in general
- **Linear supervised machine learning:** only  $n$  real numbers
  - If  $f_i(\theta) = \ell(y_i, \Phi(x_i)^\top \theta)$ , then  $f'_i(\theta) = \ell'(y_i, \Phi(x_i)^\top \theta) \Phi(x_i)$

# Stochastic average gradient - Convergence analysis

- **Assumptions**

- Each  $f_i$  is  $L$ -smooth,  $i = 1, \dots, n$  - link with  $R^2$
- $g = \frac{1}{n} \sum_{i=1}^n f_i$  is  $\mu$ -strongly convex
- constant step size  $\gamma_t = 1/(16L)$  - no need to know  $\mu$

# Stochastic average gradient - Convergence analysis

- **Assumptions**

- Each  $f_i$  is  $L$ -smooth,  $i = 1, \dots, n$  - link with  $R^2$
- $g = \frac{1}{n} \sum_{i=1}^n f_i$  is  $\mu$ -strongly convex
- constant step size  $\gamma_t = 1/(16L)$  - no need to know  $\mu$

- **Strongly convex case** (Le Roux et al., 2012, 2013)

$$\mathbb{E}[g(\theta_t) - g(\theta_*)] \leq \text{cst} \times \left(1 - \min\left\{\frac{1}{8n}, \frac{\mu}{16L}\right\}\right)^t$$

- Linear (exponential) convergence rate with  $O(d)$  iteration cost
- After one pass, reduction of cost by  $\exp\left(-\min\left\{\frac{1}{8}, \frac{n\mu}{16L}\right\}\right)$
- NB: in machine learning, may often restrict to  $\mu \geq L/n$   
 $\Rightarrow$  constant error reduction after each effective pass

# Convergence analysis - Proof sketch

- **Main step:** find “good” Lyapunov function  $J(\theta_t, y_1^t, \dots, y_n^t)$ 
  - such that  $\mathbb{E}[J(\theta_t, y_1^t, \dots, y_n^t) | \mathcal{F}_{t-1}] < J(\theta_{t-1}, y_1^{t-1}, \dots, y_n^{t-1})$
  - no natural candidates
- **Computer-aided proof**
  - Parameterize function  $J(\theta_t, y_1^t, \dots, y_n^t) = g(\theta_t) - g(\theta_*) + \text{quadratic}$
  - Solve semidefinite program to obtain candidates (that depend on  $n, \mu, L$ )
  - Check validity with symbolic computations

# Running-time comparisons (strongly-convex)

- **Assumptions:**  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$

– Each  $f_i$  convex  $L$ -smooth and  $g$   $\mu$ -strongly convex

Stochastic gradient descent	$d \times \frac{L}{\mu} \times \frac{1}{\epsilon}$
Gradient descent	$d \times n \frac{L}{\mu} \times \log \frac{1}{\epsilon}$
Accelerated gradient descent	$d \times n \sqrt{\frac{L}{\mu}} \times \log \frac{1}{\epsilon}$
SAG	$d \times \left(n + \frac{L}{\mu}\right) \times \log \frac{1}{\epsilon}$

- NB-1: for (accelerated) gradient descent,  $L =$  smoothness constant of  $g$
- NB-2: with non-uniform sampling,  $L =$  average smoothness constants of all  $f_i$ 's

# Running-time comparisons (strongly-convex)

- **Assumptions:**  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$

– Each  $f_i$  convex  $L$ -smooth and  $g$   $\mu$ -strongly convex

Stochastic gradient descent	$d \times \frac{L}{\mu} \times \frac{1}{\epsilon}$
Gradient descent	$d \times n \frac{L}{\mu} \times \log \frac{1}{\epsilon}$
Accelerated gradient descent	$d \times n \sqrt{\frac{L}{\mu}} \times \log \frac{1}{\epsilon}$
SAG	$d \times \left(n + \frac{L}{\mu}\right) \times \log \frac{1}{\epsilon}$

- **Beating two lower bounds** (Nemirovsky and Yudin, 1983; Nesterov, 2004): **with additional assumptions**

(1) stochastic gradient: exponential rate for **finite** sums

(2) full gradient: better exponential rate using the **sum structure**



# Running-time comparisons (non-strongly-convex)

- **Assumptions:**  $g(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta)$ 
  - Each  $f_i$  convex  $L$ -smooth
  - **Ill conditioned problems:**  $g$  may not be strongly-convex ( $\mu = 0$ )

Stochastic gradient descent	$d \times 1/\varepsilon^2$
Gradient descent	$d \times n/\varepsilon$
Accelerated gradient descent	$d \times n/\sqrt{\varepsilon}$
SAG	$d \times \sqrt{n}/\varepsilon$

- Adaptivity to potentially hidden strong convexity
- No need to know the local/global strong-convexity constant

# Stochastic average gradient

## Implementation details and extensions

- **Sparsity in the features**

- Just-in-time updates  $\Rightarrow$  replace  $O(d)$  by number of non zeros
- See also Leblond, Pedregosa, and Lacoste-Julien (2016)

- **Mini-batches**

- Reduces the memory requirement + block access to data

- **Line-search**

- Avoids knowing  $L$  in advance

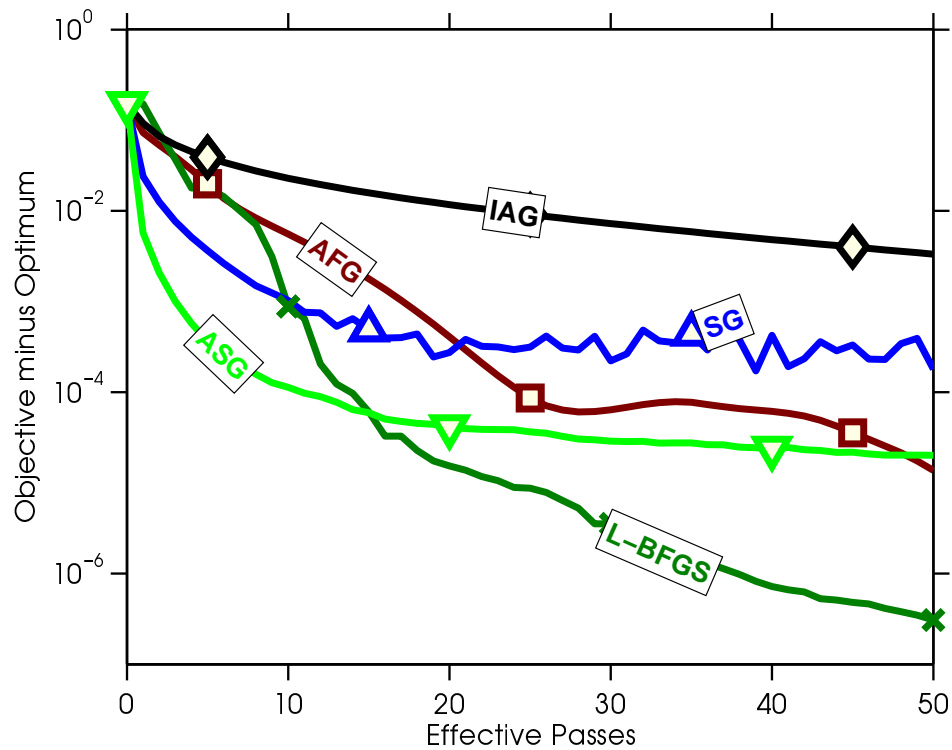
- **Non-uniform sampling**

- Favors functions with large variations

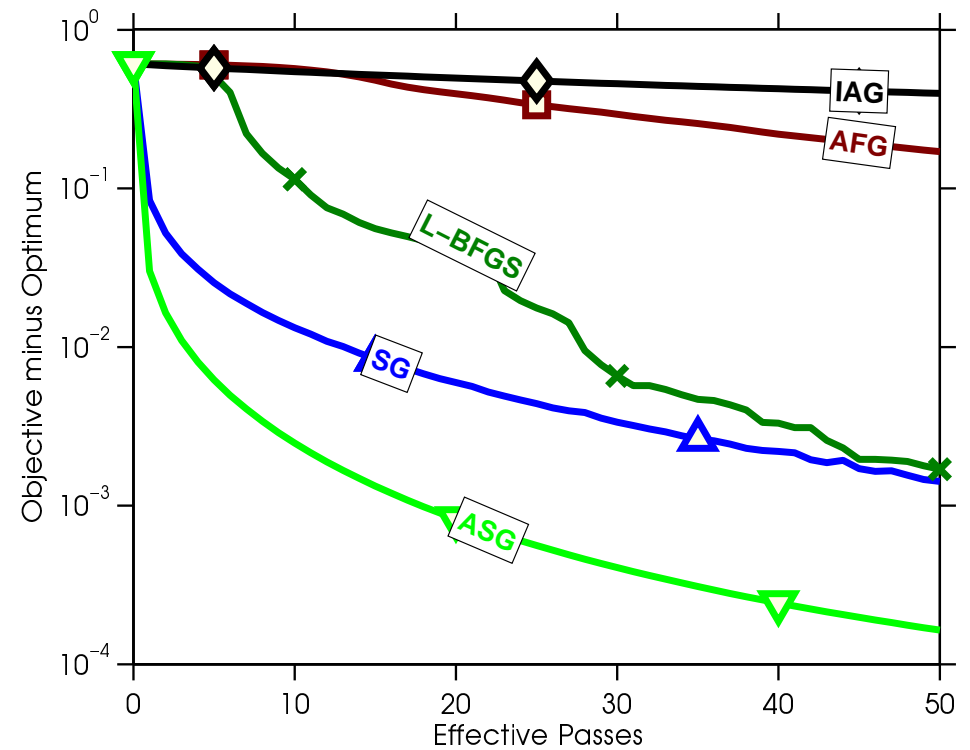
- See [www.cs.ubc.ca/~schmidtm/Software/SAG.html](http://www.cs.ubc.ca/~schmidtm/Software/SAG.html)

# Experimental results (logistic regression)

quantum dataset  
( $n = 50\,000$ ,  $d = 78$ )

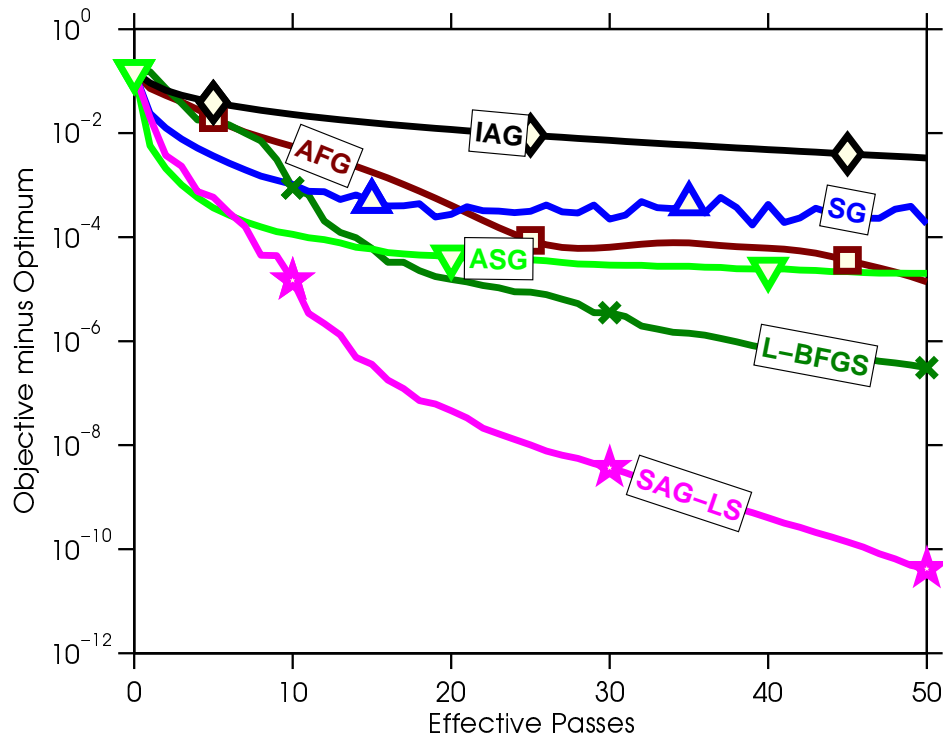


rcv1 dataset  
( $n = 697\,641$ ,  $d = 47\,236$ )

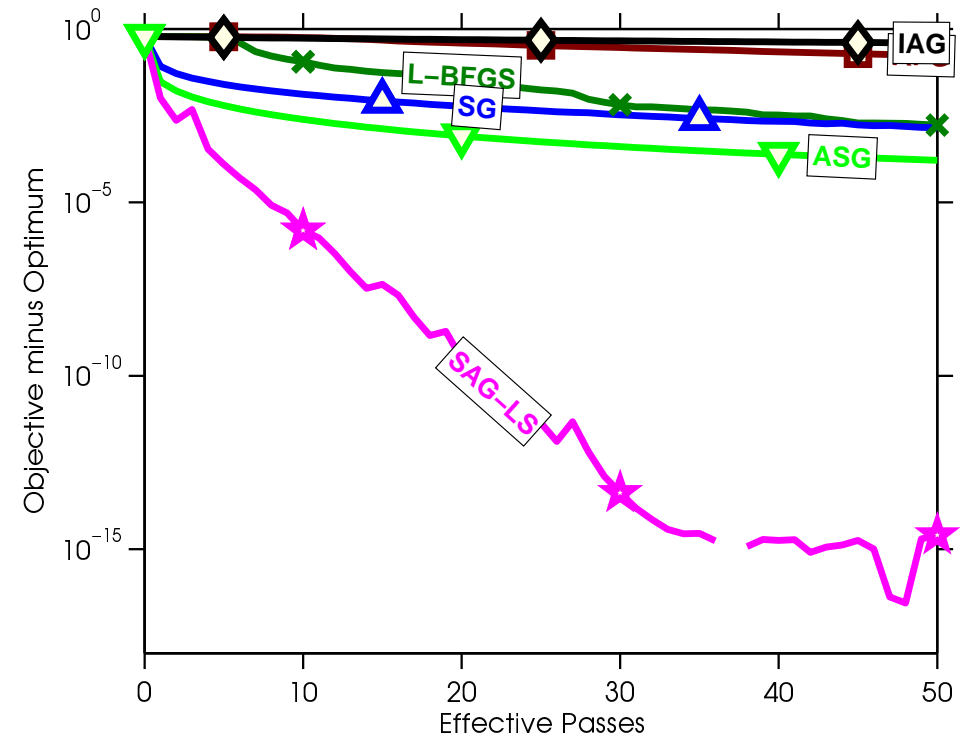


# Experimental results (logistic regression)

quantum dataset  
( $n = 50\,000$ ,  $d = 78$ )

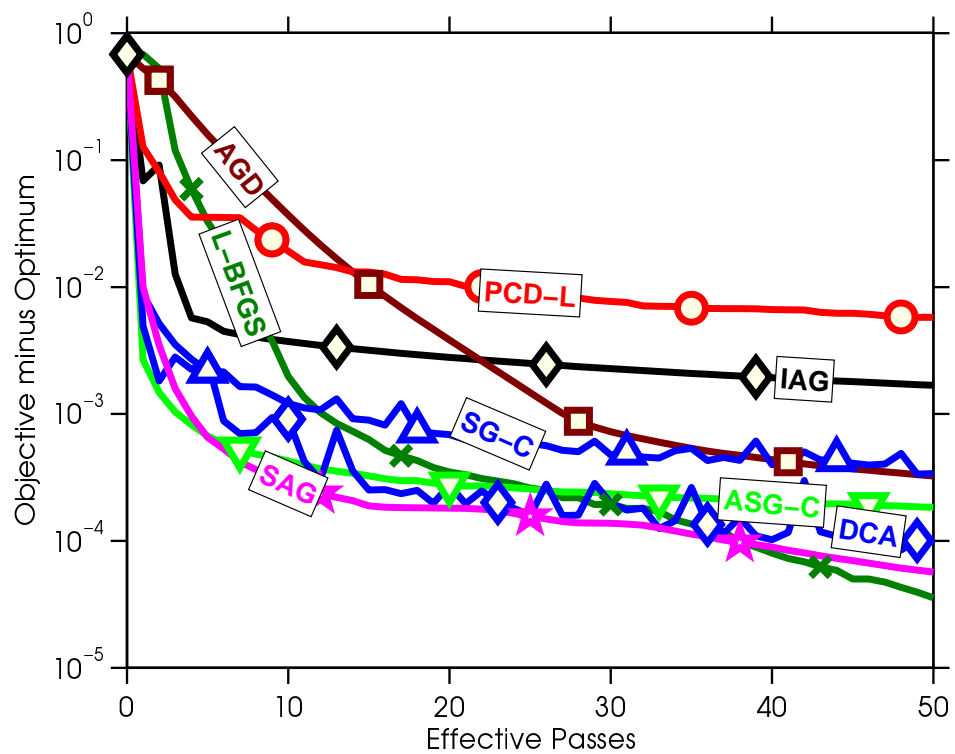


rcv1 dataset  
( $n = 697\,641$ ,  $d = 47\,236$ )

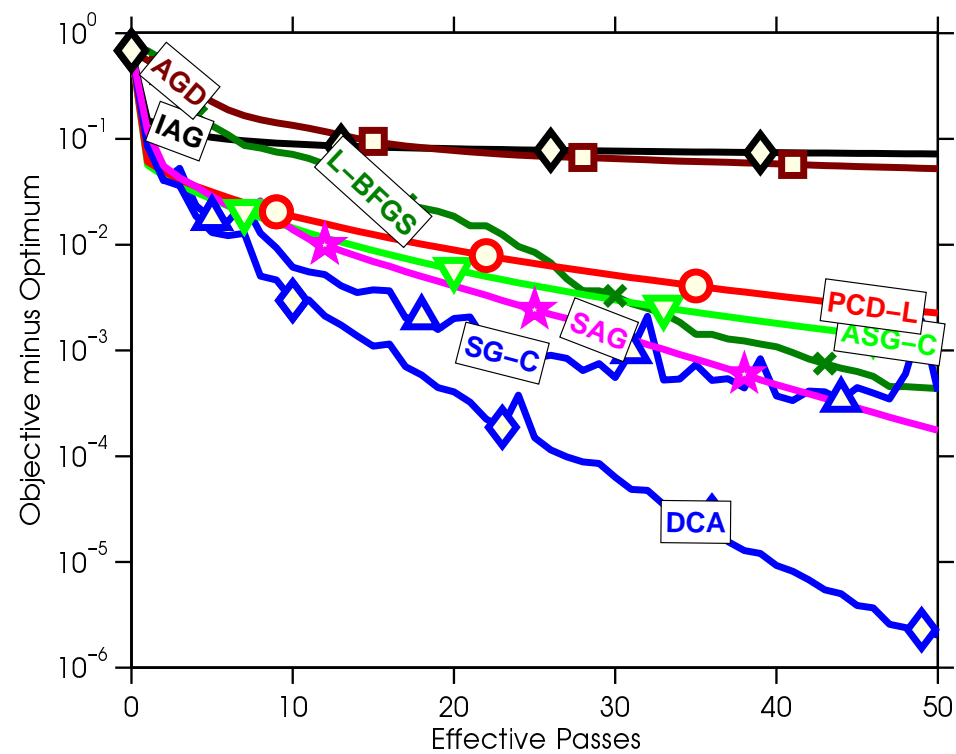


# Before non-uniform sampling

protein dataset  
( $n = 145\,751$ ,  $d = 74$ )

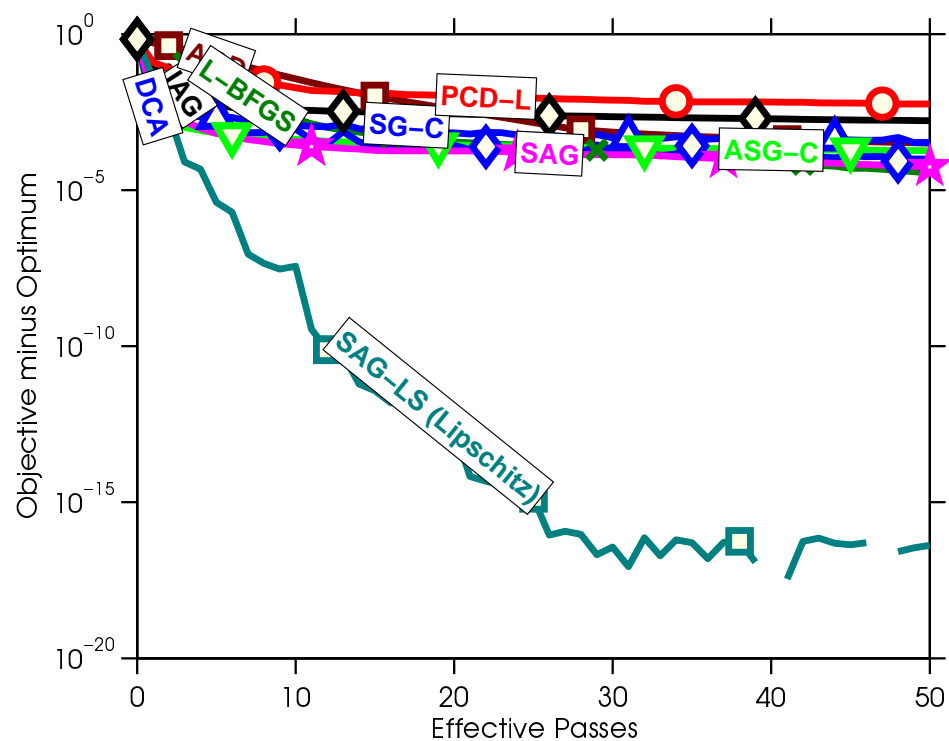


sido dataset  
( $n = 12\,678$ ,  $d = 4\,932$ )

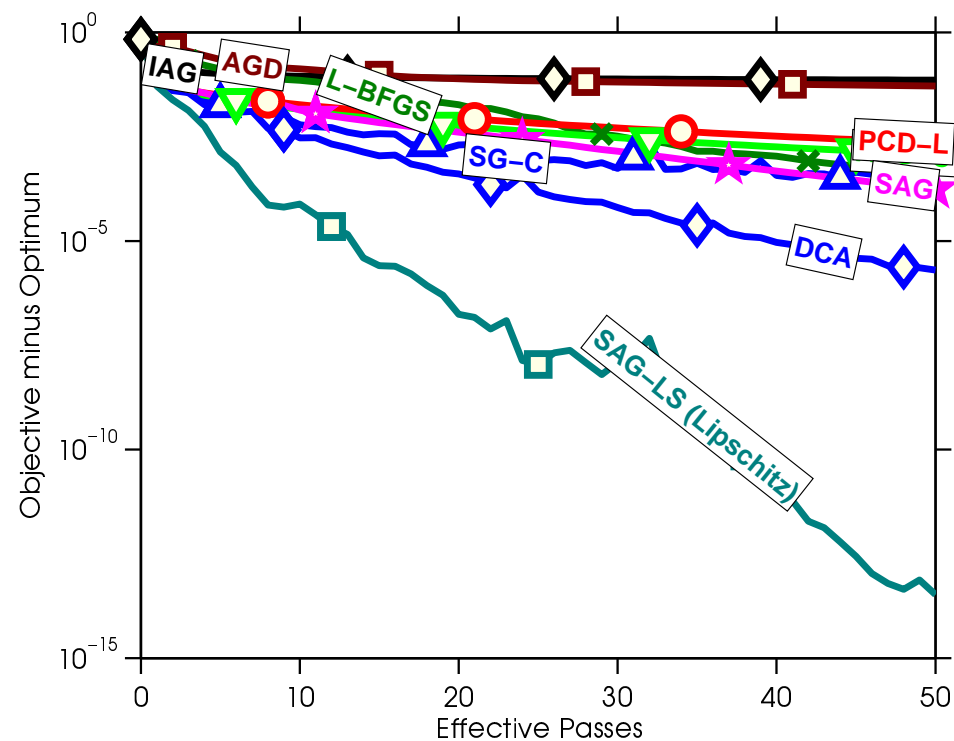


# After non-uniform sampling

protein dataset  
( $n = 145\,751$ ,  $d = 74$ )



sido dataset  
( $n = 12\,678$ ,  $d = 4\,932$ )



# Linearly convergent stochastic gradient algorithms

- **Many related algorithms**
  - SAG (Le Roux, Schmidt, and Bach, 2012)
  - SDCA (Shalev-Shwartz and Zhang, 2012)
  - SVRG (Johnson and Zhang, 2013; Zhang et al., 2013)
  - MISO (Mairal, 2015)
  - Finito (Defazio et al., 2014a)
  - SAGA (Defazio, Bach, and Lacoste-Julien, 2014b)
  - ...
- **Similar rates of convergence and iterations**

# Linearly convergent stochastic gradient algorithms

- **Many related algorithms**
  - SAG (Le Roux, Schmidt, and Bach, 2012)
  - SDCA (Shalev-Shwartz and Zhang, 2012)
  - SVRG (Johnson and Zhang, 2013; Zhang et al., 2013)
  - MISO (Mairal, 2015)
  - Finito (Defazio et al., 2014a)
  - SAGA (Defazio, Bach, and Lacoste-Julien, 2014b)
  - ...
- **Similar rates of convergence and iterations**
- **Different interpretations and proofs / proof lengths**
  - Lazy gradient evaluations
  - Variance reduction



# Variance reduction

- **Principle:** reducing variance of sample of  $X$  by using a sample from another random variable  $Y$  with known expectation

$$Z_\alpha = \alpha(X - Y) + \mathbb{E}Y$$

- $\mathbb{E}Z_\alpha = \alpha\mathbb{E}X + (1 - \alpha)\mathbb{E}Y$
- $\text{var}(Z_\alpha) = \alpha^2 [\text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)]$
- $\alpha = 1$ : no bias,  $\alpha < 1$ : potential bias (but reduced variance)
- Useful if  $Y$  positively correlated with  $X$

# Variance reduction

- **Principle:** reducing variance of sample of  $X$  by using a sample from another random variable  $Y$  with known expectation

$$Z_\alpha = \alpha(X - Y) + \mathbb{E}Y$$

- $\mathbb{E}Z_\alpha = \alpha\mathbb{E}X + (1 - \alpha)\mathbb{E}Y$
  - $\text{var}(Z_\alpha) = \alpha^2 [\text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)]$
  - $\alpha = 1$ : no bias,  $\alpha < 1$ : potential bias (but reduced variance)
  - Useful if  $Y$  positively correlated with  $X$
- **Application to gradient estimation** (Johnson and Zhang, 2013; Zhang, Mahdavi, and Jin, 2013)
    - SVRG:  $X = f'_{i(t)}(\theta_{t-1})$ ,  $Y = f'_{i(t)}(\tilde{\theta})$ ,  $\alpha = 1$ , with  $\tilde{\theta}$  stored
    - $\mathbb{E}Y = \frac{1}{n} \sum_{i=1}^n f'_i(\tilde{\theta})$  full gradient at  $\tilde{\theta}$ ,  $X - Y = f'_{i(t)}(\theta_{t-1}) - f'_{i(t)}(\tilde{\theta})$

# Stochastic variance reduced gradient (SVRG) (Johnson and Zhang, 2013; Zhang et al., 2013)

- Initialize  $\tilde{\theta} \in \mathbb{R}^d$
- For  $i_{\text{epoch}} = 1$  to  $\#$  of epochs
  - Compute all gradients  $f'_i(\tilde{\theta})$  ; store  $g'(\tilde{\theta}) = \frac{1}{n} \sum_{i=1}^n f'_i(\tilde{\theta})$
  - Initialize  $\theta_0 = \tilde{\theta}$
  - For  $t = 1$  to **length of epochs**
    - $$\theta_t = \theta_{t-1} - \gamma \left[ g'(\tilde{\theta}) + (f'_{i(t)}(\theta_{t-1}) - f'_{i(t)}(\tilde{\theta})) \right]$$
  - Update  $\tilde{\theta} = \theta_t$
- Output:  $\tilde{\theta}$

# Stochastic variance reduced gradient (SVRG) (Johnson and Zhang, 2013; Zhang et al., 2013)

- Initialize  $\tilde{\theta} \in \mathbb{R}^d$
- For  $i_{\text{epoch}} = 1$  to  $\#$  of epochs
  - Compute all gradients  $f'_i(\tilde{\theta})$  ; store  $g'(\tilde{\theta}) = \frac{1}{n} \sum_{i=1}^n f'_i(\tilde{\theta})$
  - Initialize  $\theta_0 = \tilde{\theta}$
  - For  $t = 1$  to **length of epochs**
    - $$\theta_t = \theta_{t-1} - \gamma \left[ g'(\tilde{\theta}) + (f'_{i(t)}(\theta_{t-1}) - f'_{i(t)}(\tilde{\theta})) \right]$$
  - Update  $\tilde{\theta} = \theta_t$
- Output:  $\tilde{\theta}$

- **No need to store gradients** - two gradient evaluations per inner step
- Two parameters: length of epochs + step-size  $\gamma$
- Same linear convergence rate as SAG, simpler proof

# Stochastic variance reduced gradient (SVRG)

- **Algorithm divide into “epochs”**
- At each epoch, starting from  $\theta_0 = \tilde{\theta}$ , perform the iteration
  - Sample  $i_t$  uniformly at random
  - Gradient step:  $\theta_t = \theta_{t-1} - \gamma \left[ f'_{i_t}(\theta_{t-1}) - f'_{i_t}(\tilde{\theta}) + g'(\tilde{\theta}) \right]$
- **Proposition:** If each  $f_i$  is  $R^2$ -smooth and  $g = \frac{1}{n} \sum_{i=1}^n f_i$  is  $\mu$ -strongly convex, then after  $k = 20R^2/\mu$  steps and with  $\gamma = 1/10R^2$ , then  $f(\theta) - f(\theta_*)$  is reduced by 10%

## SVRG proof - from Bubeck (2015)

• **Lemma:**  $\mathbb{E}\|f'_i(\theta) - f'_i(\theta_*)\|^2 \leq 2R^2 [g(\theta) - g(\theta_*)]$

– Proof:  $\mathbb{E}\|f'_i(\theta) - f'_i(\theta_*)\|^2 \leq 2R^2 \mathbb{E} [f_i(\theta) - f_i(\theta_*) - f'_i(\theta_*)^\top (\theta - \theta_*)]$   
by the proof of co-coercivity, which is equal to  $2R^2 [g'(\theta) - g(\theta_*)]$

# SVRG proof - from Bubeck (2015)

- **Lemma:**  $\mathbb{E}\|f'_i(\theta) - f'_i(\theta_*)\|^2 \leq 2R^2 [g(\theta) - g(\theta_*)]$
- From iteration  $\theta_t = \theta_{t-1} - \gamma [f'_{i_t}(\theta_{t-1}) - f'_{i_t}(\tilde{\theta}) + g'(\tilde{\theta})] = \theta_{t-1} - \gamma g_t$

$$\begin{aligned}
 \|\theta_t - \theta_*\|^2 &= \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g_t + \gamma^2 \|g_t\|^2 \\
 \mathbb{E}[\|\theta_t - \theta_*\|^2 | \mathcal{F}_{t-1}] &\leq \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\
 &\quad + 2\gamma^2 \|f'_{i_t}(\theta_{t-1}) - f'_{i_t}(\theta_*)\|^2 + 2\gamma^2 \|f'_{i_t}(\tilde{\theta}) - f'_{i_t}(\theta_*) - g'(\tilde{\theta})\|^2 \\
 &\leq \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\
 &\quad + 2\gamma^2 R^2 [g(\theta_{t-1}) - g(\theta_*) + g(\tilde{\theta}) - g(\theta_*)] \\
 &\leq \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(1 - 2\gamma R^2)[g(\theta_{t-1}) - g(\theta_*)] + 4R^2\gamma^2[g(\tilde{\theta}) - g(\theta_*)]
 \end{aligned}$$

- By summing  $k$  times, we get:

$$\mathbb{E}\|\theta_k - \theta_*\|^2 \leq \|\theta_0 - \theta_*\|^2 - 2\gamma(1 - 2\gamma R^2) \sum_{t=1}^k \mathbb{E}[g(\theta_{t-1}) - g(\theta_*)] + 4kR^2\gamma^2[g(\tilde{\theta}) - g(\theta_*)]$$

which leads to the desired result

# Interpretation of SAG as variance reduction

- **SAG update:**  $\theta_t = \theta_{t-1} - \frac{\gamma}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$ 
  - Interpretation as lazy gradient evaluations



# Interpretation of SAG as variance reduction

- **SAG update:**  $\theta_t = \theta_{t-1} - \frac{\gamma}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

- Interpretation as lazy gradient evaluations

- **SAG update:**  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n y_i^{t-1} + \frac{1}{n} (f'_{i(t)}(\theta_{t-1}) - y_{i(t)}^{t-1}) \right]$

- Biased update (expectation w.r.t. to  $i(t)$  not equal to full gradient)

# Interpretation of SAG as variance reduction

- **SAG update:**  $\theta_t = \theta_{t-1} - \frac{\gamma}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

- Interpretation as lazy gradient evaluations

- **SAG update:**  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n y_i^{t-1} + \frac{1}{n} (f'_{i(t)}(\theta_{t-1}) - y_{i(t)}^{t-1}) \right]$

- Biased update (expectation w.r.t. to  $i(t)$  not equal to full gradient)

- **SVRG update:**  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n f'_i(\tilde{\theta}) + (f'_{i(t)}(\theta_{t-1}) - f'_{i(t)}(\tilde{\theta})) \right]$

- Unbiased update

# Interpretation of SAG as variance reduction

- **SAG update:**  $\theta_t = \theta_{t-1} - \frac{\gamma}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$ 
  - Interpretation as lazy gradient evaluations
- **SAG update:**  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n y_i^{t-1} + \frac{1}{n} (f'_{i(t)}(\theta_{t-1}) - y_{i(t)}^{t-1}) \right]$ 
  - Biased update (expectation w.r.t. to  $i(t)$  not equal to full gradient)
- **SVRG update:**  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n f'_i(\tilde{\theta}) + (f'_{i(t)}(\theta_{t-1}) - f'_{i(t)}(\tilde{\theta})) \right]$ 
  - Unbiased update
- **SAGA update:**  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n y_i^{t-1} + (f'_{i(t)}(\theta_{t-1}) - y_{i(t)}^{t-1}) \right]$ 
  - Defazio, Bach, and Lacoste-Julien (2014b)
  - Unbiased update without epochs

# SVRG vs. SAGA

- **SAGA** update:  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n y_i^{t-1} + (f'_{i(t)}(\theta_{t-1}) - y_{i(t)}^{t-1}) \right]$
- **SVRG** update:  $\theta_t = \theta_{t-1} - \gamma \left[ \frac{1}{n} \sum_{i=1}^n f'_i(\tilde{\theta}) + (f'_{i(t)}(\theta_{t-1}) - f'_{i(t)}(\tilde{\theta})) \right]$

	SAGA	SVRG
<b>Storage of gradients</b>	<b>yes</b>	<b>no</b>
Epoch-based	no	yes
Parameters	step-size	step-size & epoch lengths
Gradient evaluations per step	1	at least 2
Adaptivity to strong-convexity	yes	no
Robustness to ill-conditioning	yes	no

– See Babanezhad et al. (2015)

# Proximal extensions

- **Composite** optimization problems:  $\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(\theta) + h(\theta)$ 
  - $f_i$  smooth and convex
  - $h$  convex, potentially non-smooth

# Proximal extensions

- **Composite optimization problems:**  $\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(\theta) + h(\theta)$ 
  - $f_i$  smooth and convex
  - $h$  convex, potentially non-smooth
  - Constrained optimization:  $h(\theta) = 0$  if  $\theta \in K$ , and  $+\infty$  otherwise
  - Sparsity-inducing norms, e.g.,  $h(\theta) = \|\theta\|_1$

# Proximal extensions

- **Composite optimization problems:**  $\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(\theta) + h(\theta)$

- $f_i$  smooth and convex
- $h$  convex, potentially non-smooth
- **Constrained optimization:**  $h(\theta) = 0$  if  $\theta \in K$ , and  $+\infty$  otherwise
- **Sparsity-inducing norms**, e.g.,  $h(\theta) = \|\theta\|_1$

- **Proximal methods (a.k.a. splitting methods)**

- Extra projection / soft thresholding step after gradient update
- See, e.g., Combettes and Pesquet (2011); Bach, Jenatton, Mairal, and Obozinski (2012b); Parikh and Boyd (2014)

# Proximal extensions

- **Composite optimization problems:**  $\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(\theta) + h(\theta)$ 
  - $f_i$  smooth and convex
  - $h$  convex, potentially non-smooth
  - Constrained optimization:  $h(\theta) = 0$  if  $\theta \in K$ , and  $+\infty$  otherwise
  - Sparsity-inducing norms, e.g.,  $h(\theta) = \|\theta\|_1$
- **Proximal methods (a.k.a. splitting methods)**
  - Extra projection / soft thresholding step after gradient update
  - See, e.g., Combettes and Pesquet (2011); Bach, Jenatton, Mairal, and Obozinski (2012b); Parikh and Boyd (2014)
- **Directly extends to variance-reduced gradient techniques**
  - Same rates of convergence



# Acceleration

- **Similar guarantees for finite sums:** SAG, SDCA, SVRG (Xiao and Zhang, 2014), SAGA, MISO (Mairal, 2015)

Gradient descent	$d \times$	$n \frac{L}{\mu}$	$\times \log \frac{1}{\epsilon}$
Accelerated gradient descent	$d \times$	$n \sqrt{\frac{L}{\mu}}$	$\times \log \frac{1}{\epsilon}$
SAG(A), SVRG, SDCA, MISO	$d \times$	$(n + \frac{L}{\mu})$	$\times \log \frac{1}{\epsilon}$

# Acceleration

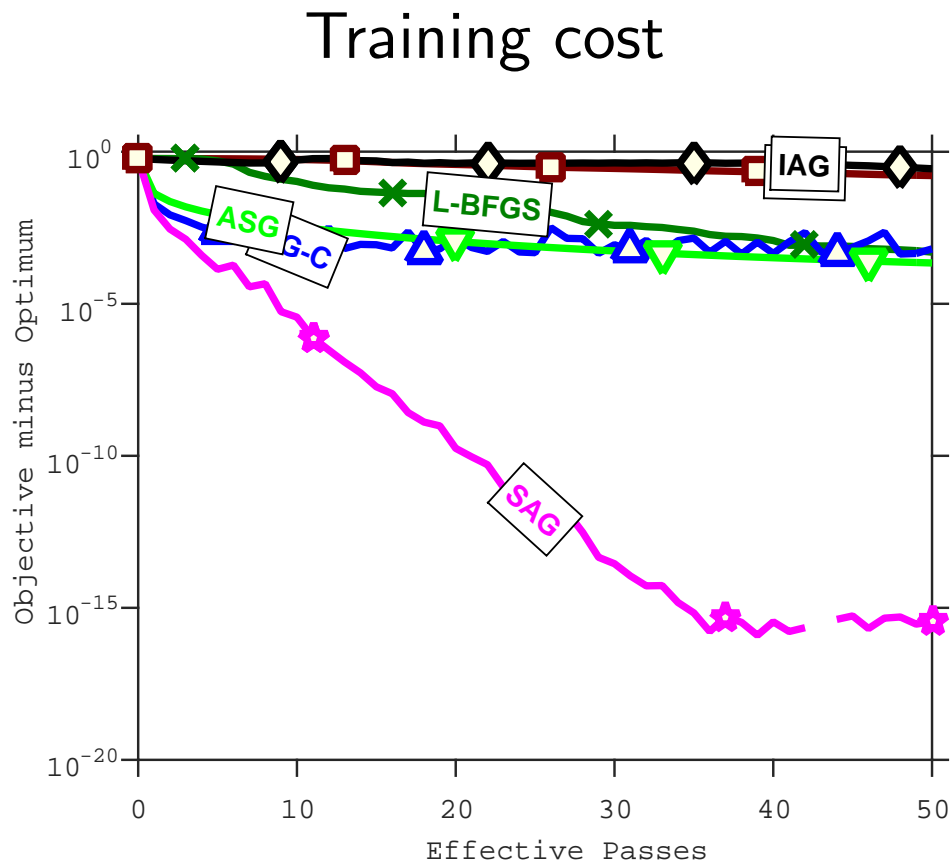
- **Similar guarantees for finite sums:** SAG, SDCA, SVRG (Xiao and Zhang, 2014), SAGA, MISO (Mairal, 2015)

Gradient descent	$d \times n \frac{L}{\mu} \times \log \frac{1}{\epsilon}$
Accelerated gradient descent	$d \times n \sqrt{\frac{L}{\mu}} \times \log \frac{1}{\epsilon}$
SAG(A), SVRG, SDCA, MISO	$d \times (n + \frac{L}{\mu}) \times \log \frac{1}{\epsilon}$
<b>Accelerated versions</b>	$d \times (n + \sqrt{n \frac{L}{\mu}}) \times \log \frac{1}{\epsilon}$

- **Acceleration for special algorithms** (e.g., Shalev-Shwartz and Zhang, 2014; Nitanda, 2014; Lan, 2015)
- **Catalyst** (Lin, Mairal, and Harchaoui, 2015)
  - Widely applicable generic acceleration scheme

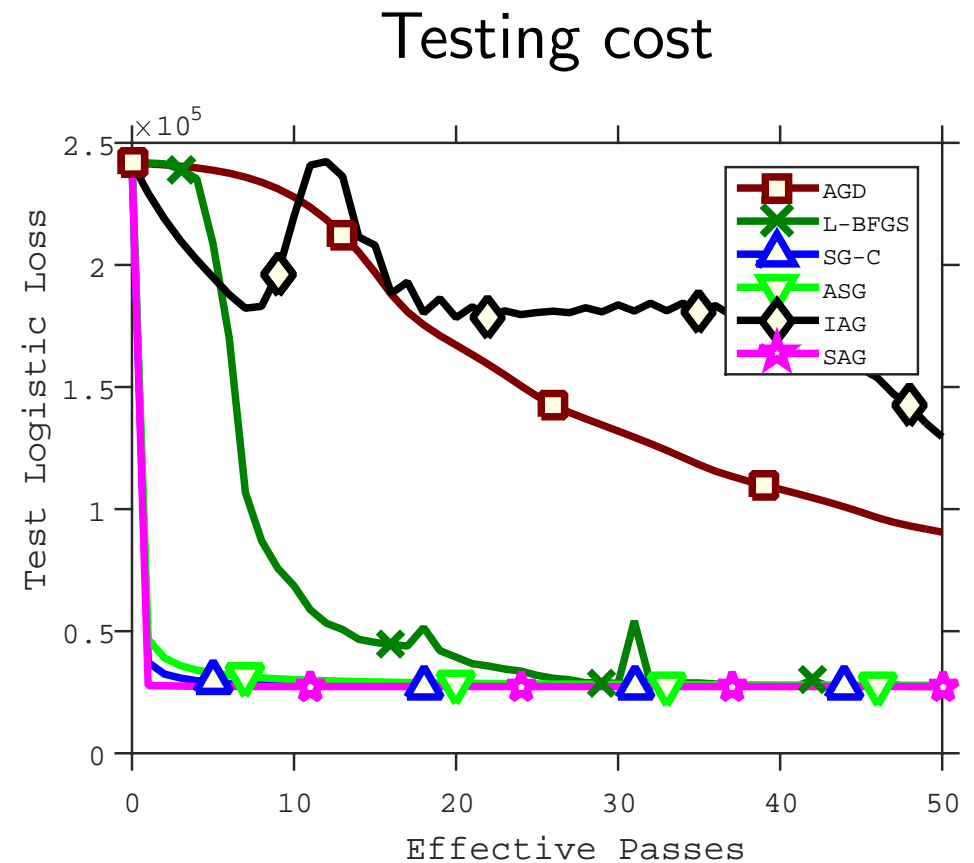
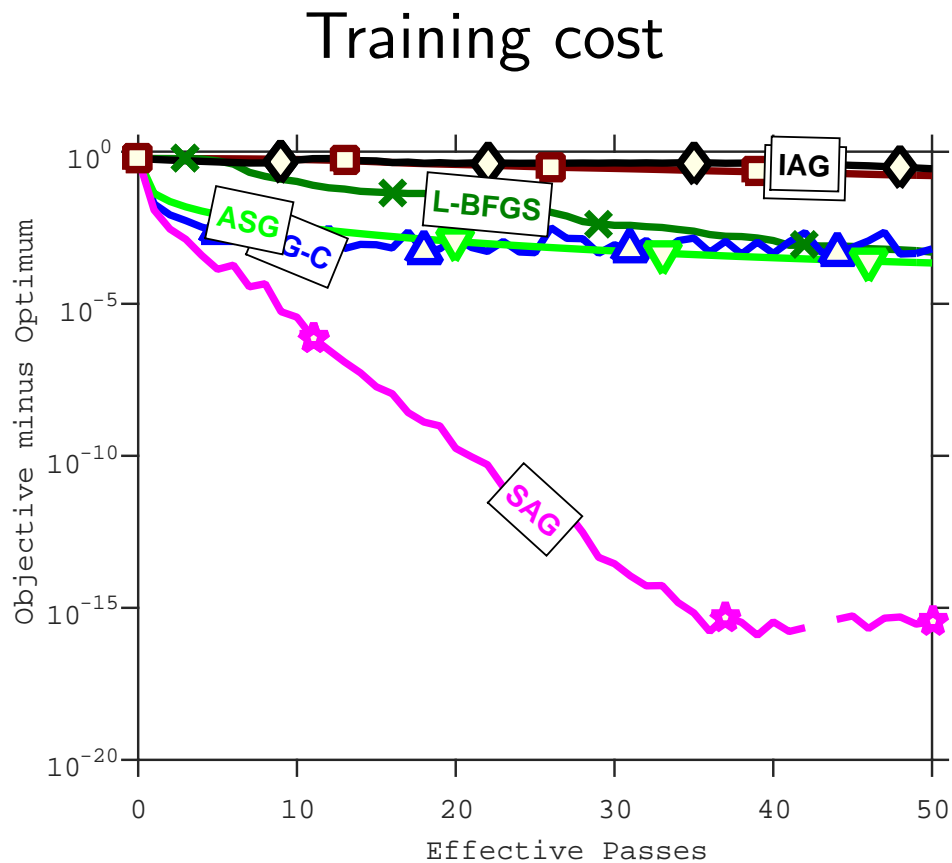
# From training to testing errors

- rcv1 dataset ( $n = 697\,641$ ,  $d = 47\,236$ )
  - NB: IAG, SG-C, ASG with optimal step-sizes in hindsight



# From training to testing errors

- rcv1 dataset ( $n = 697\,641$ ,  $d = 47\,236$ )
  - NB: IAG, SG-C, ASG with optimal step-sizes in hindsight



# SGD minimizes the testing cost!

- **Goal:** minimize  $f(\theta) = \mathbb{E}_{p(x,y)} \ell(y, \theta^\top \Phi(x))$ 
  - Given  $n$  independent samples  $(x_i, y_i)$ ,  $i = 1, \dots, n$  from  $p(x, y)$
  - Given a **single pass** of stochastic gradient descent
  - Bounds on the excess **testing** cost  $\mathbb{E} f(\bar{\theta}_n) - \inf_{\theta \in \mathbb{R}^d} f(\theta)$

# SGD minimizes the testing cost!

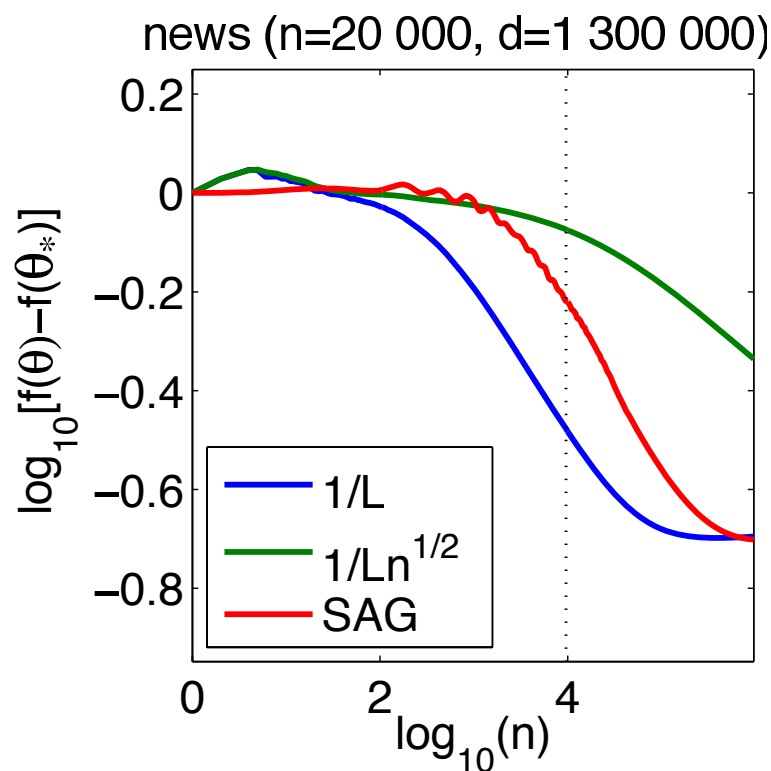
- **Goal:** minimize  $f(\theta) = \mathbb{E}_{p(x,y)} \ell(y, \theta^\top \Phi(x))$ 
  - Given  $n$  independent samples  $(x_i, y_i)$ ,  $i = 1, \dots, n$  from  $p(x, y)$
  - Given a **single pass** of stochastic gradient descent
  - Bounds on the excess **testing** cost  $\mathbb{E}f(\bar{\theta}_n) - \inf_{\theta \in \mathbb{R}^d} f(\theta)$
- **Optimal convergence rates:**  $O(1/\sqrt{n})$  and  $O(1/(n\mu))$ 
  - Optimal for non-smooth losses (Nemirovsky and Yudin, 1983)
  - Attained by averaged SGD with decaying step-sizes

# SGD minimizes the testing cost!

- **Goal:** minimize  $f(\theta) = \mathbb{E}_{p(x,y)} \ell(y, \theta^\top \Phi(x))$ 
  - Given  $n$  independent samples  $(x_i, y_i)$ ,  $i = 1, \dots, n$  from  $p(x, y)$
  - Given a **single pass** of stochastic gradient descent
  - Bounds on the excess **testing** cost  $\mathbb{E}f(\bar{\theta}_n) - \inf_{\theta \in \mathbb{R}^d} f(\theta)$
- **Optimal convergence rates:**  $O(1/\sqrt{n})$  and  $O(1/(n\mu))$ 
  - Optimal for non-smooth losses (Nemirovsky and Yudin, 1983)
  - Attained by averaged SGD with decaying step-sizes
- **Constant-step-size SGD**
  - Linear convergence up to the noise level for strongly-convex problems (Solodov, 1998; Nedic and Bertsekas, 2000)
  - **Full convergence and robustness to ill-conditioning?**

# Robust averaged stochastic gradient (Bach and Moulines, 2013)

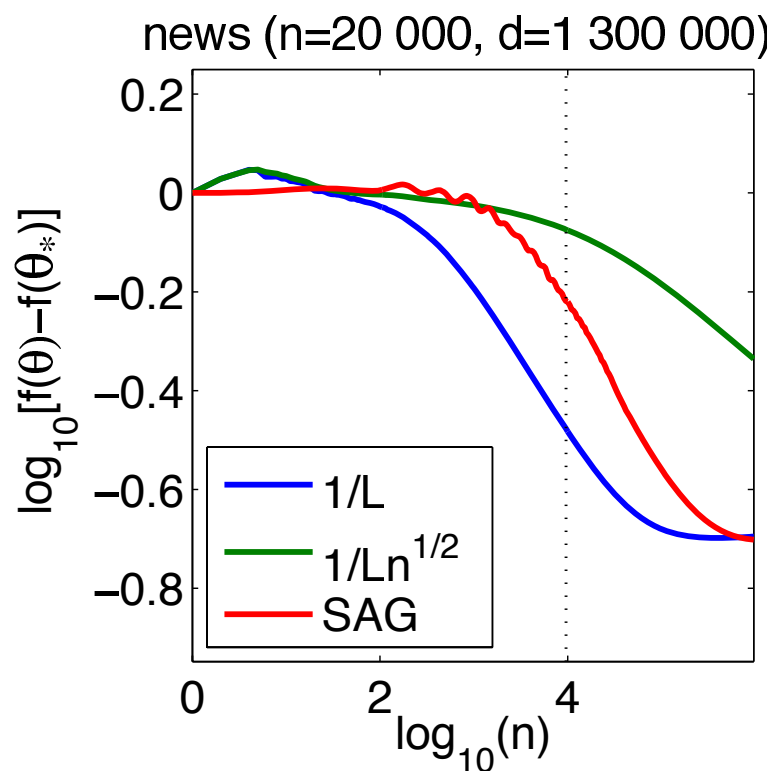
- **Constant-step-size SGD is convergent for least-squares**
  - Convergence rate in  $O(1/n)$  without any dependence on  $\mu$
  - Simple choice of step-size (equal to  $1/L$ )





# Robust averaged stochastic gradient (Bach and Moulines, 2013)

- Constant-step-size SGD is convergent for least-squares
  - Convergence rate in  $O(1/n)$  without any dependence on  $\mu$
  - Simple choice of step-size (equal to  $1/L$ )



- Convergence in  $O(1/n)$  for smooth losses with  $O(d)$  online Newton step

# Conclusions - variance reduction

- **Linearly-convergent stochastic gradient methods**
  - Provable and precise rates
  - Improves on two known lower-bounds (by using structure)
  - Several extensions / interpretations / accelerations

# Conclusions - variance reduction

- **Linearly-convergent stochastic gradient methods**
  - Provable and precise rates
  - Improves on two known lower-bounds (by using structure)
  - Several extensions / interpretations / accelerations
- **Extensions and future work**
  - Extension to saddle-point problems (Balamurugan and Bach, 2016)
  - Lower bounds for finite sums (Agarwal and Bottou, 2014; Lan, 2015; Arjevani and Shamir, 2016)
  - Sampling without replacement (Gurbuzbalaban et al., 2015; Shamir, 2016)

# Conclusions - variance reduction

- **Linearly-convergent stochastic gradient methods**
  - Provable and precise rates
  - Improves on two known lower-bounds (by using structure)
  - Several extensions / interpretations / accelerations
- **Extensions and future work**
  - Extension to saddle-point problems (Balamurugan and Bach, 2016)
  - Lower bounds for finite sums (Agarwal and Bottou, 2014; Lan, 2015; Arjevani and Shamir, 2016)
  - Sampling without replacement (Gurbuzbalaban et al., 2015; Shamir, 2016)
  - Bounds on testing errors for incremental methods (Frostig et al., 2015; Babanezhad et al., 2015)

# Fundamentals of constrained optimization

- We consider the following **primal** optimization problem

$$\min_{x \in D} f(x) \quad \text{s.t.} \quad \forall i \in \{1, \dots, m\}, h_i(x) = 0 \quad \text{and} \quad \forall j \in \{1, \dots, r\}, g_j(x) \leq 0$$

- We denote by  $D^*$  the set of  $x \in D$  satisfying the constraints

# Fundamentals of constrained optimization

- We consider the following **primal** optimization problem

$$\min_{x \in D} f(x) \quad \text{s.t.} \quad \forall i \in \{1, \dots, m\}, h_i(x) = 0 \quad \text{and} \quad \forall j \in \{1, \dots, r\}, g_j(x) \leq 0$$

- We denote by  $D^*$  the set of  $x \in D$  satisfying the constraints
- **Lagrangian:** function  $\mathcal{L} : \mathbb{R}^m \times \mathbb{R}_+^r$  defined as

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)$$

- $\lambda$  et  $\mu$  are called Lagrange multipliers or dual variables
- Primal problem = supremum of Lagrangian with respect to dual variables: for all  $x \in D$ ,

$$\sup_{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r} \mathcal{L}(x, \lambda, \mu) = \begin{cases} f(x) & \text{si } x \in D^* \\ +\infty & \text{otherwise} \end{cases}$$

# Fundamentals of constrained optimization

- **Primal problem** equivalent to  $p^* = \inf_{x \in D} \sup_{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r} \mathcal{L}(x, \lambda, \mu)$
- **Dual function:**  $q(\lambda, \mu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \mu) = \inf_{x \in D} f(x) + \lambda^\top h(x) + \mu^\top g(x)$
- **Dual problem:** minimization of  $q$  on  $\mathbb{R}^m \times \mathbb{R}_+^r$ , equivalent to
$$d^* = \sup_{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r} \inf_{x \in D} \mathcal{L}(x, \lambda, \mu).$$
  - Concave maximization problem (no assumption)

# Fundamentals of constrained optimization

- **Primal problem** equivalent to  $p^* = \inf_{x \in D} \sup_{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r} \mathcal{L}(x, \lambda, \mu)$
- **Dual function:**  $q(\lambda, \mu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \mu) = \inf_{x \in D} f(x) + \lambda^\top h(x) + \mu^\top g(x)$
- **Dual problem:** minimization of  $q$  on  $\mathbb{R}^m \times \mathbb{R}_+^r$ , equivalent to
$$d^* = \sup_{(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r} \inf_{x \in D} \mathcal{L}(x, \lambda, \mu).$$
  - Concave maximization problem (no assumption)
- **Weak duality** (no assumption):  $\forall (\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r, \forall x \in D^*$

$$\inf_{x' \in D} \mathcal{L}(x', \lambda, \mu) \leq \mathcal{L}(x, \lambda, \mu) \leq \sup_{(\lambda', \mu') \in \mathbb{R}^m \times \mathbb{R}_+^r} \mathcal{L}(x, \lambda', \mu')$$

which implies  $q(\lambda, \mu) \leq f(x)$  and thus  $d^* \leq p^*$



# Sufficient conditions for strong duality

- **Geometric interpretation** for  $\min_{x \in D} f(x)$  s.t.  $g(x) \leq 0$ 
  - Consider  $A = \{(u, t) \in \mathbb{R}^2, \exists x \in D, f(x) \leq t, g(x) \leq u\}$
- **Slater's conditions**
  - $D$  is convex,  $h_i$  affine and  $g_j$  convex and there is a strictly feasible point, that is  $\exists \bar{x} \in D^*$  such that  $\forall j, g_j(\bar{x}) < 0$
  - then  $d^* = p^*$  (**strong duality**).

# Sufficient conditions for strong duality

- **Geometric interpretation** for  $\min_{x \in D} f(x)$  s.t.  $g(x) \leq 0$ 
  - Consider  $A = \{(u, t) \in \mathbb{R}^2, \exists x \in D, f(x) \leq t, g(x) \leq u\}$
- **Slater's conditions**
  - $D$  is convex,  $h_i$  affine and  $g_j$  convex and there is a strictly feasible point, that is  $\exists \bar{x} \in D^*$  such that  $\forall j, g_j(\bar{x}) < 0$
  - then  $d^* = p^*$  (**strong duality**).
- **Karush-Kuhn-Tucker (KKT) conditions:** If strong duality holds, then  $x^*$  is primal optimal and  $(\lambda^*, \mu^*)$  are dual optimal **if and only if:**
  - *Primal stationarity:*  $x^*$  minimizes  $x \mapsto \mathcal{L}(x, \lambda^*, \mu^*)$ .
  - *Feasibility:*  $x^*$  and  $(\lambda^*, \mu^*)$  are feasible
  - *Complementary slackness:*  $\forall j, \mu_j^* g_j(x^*) = 0$

# Strong duality: remarks and examples

- **Remarks:** (a) the dual of the dual is the primal, (b) potentially several dual problems, (c) strong duality does not always hold

- **Linear programming:**  $\min_{Ax=b, x \geq 0} c^\top x = \max_{A^\top y \leq c} b^\top y$

- **Quadratic programming with equality constraint:**

$$\min_{a^\top x = b} \frac{1}{2} x^\top Q x - q^\top x$$

- **Lagrangian relaxation for combinatorial problem - Max Cut:**

$$\min_{x \in \{-1, 1\}^n} x^\top W x$$

- **Strong duality for non convex problem:**  $\min_{x^\top x \leq 1} \frac{1}{2} x^\top Q x - q^\top x$

# Dual stochastic coordinate ascent - I

- General learning formulation:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell_i(\theta^\top \Phi(x_i)) + \frac{\mu}{2} \|\theta\|_2^2$$

$$= \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 \text{ such that } \forall i, u_i = \theta^\top \Phi(x_i)$$

$$= \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^n \alpha_i (u_i - \theta^\top \Phi(x_i))$$

$$= \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^n \alpha_i (u_i - \theta^\top \Phi(x_i))$$

$$= \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^n \alpha_i (u_i - \theta^\top \Phi(x_i))$$

# Dual stochastic coordinate ascent - II

- General learning formulation:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell_i(\theta^\top \Phi(x_i)) + \frac{\mu}{2} \|\theta\|_2^2$$

$$= \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^d, \mathbf{u} \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^n \alpha_i (u_i - \theta^\top \Phi(x_i))$$

$$= \max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \min_{u_i \in \mathbb{R}} \left\{ \frac{1}{n} \ell_i(u_i) + \alpha_i u_i \right\} - \frac{1}{2\mu} \left\| \sum_{i=1}^n \alpha_i \Phi(x_i) \right\|_2^2$$

$$= \max_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^n \psi_i(\alpha_i) - \frac{1}{2\mu} \left\| \sum_{i=1}^n \alpha_i \Phi(x_i) \right\|_2^2$$

- Minimizers obtained as  $\theta = \frac{1}{\mu} \sum_{i=1}^n \alpha_i \Phi(x_i)$
- $\psi_i$  convex (up to affine transform = Fenchel-Legendre dual of  $\ell_i$ )

# Dual stochastic coordinate ascent - III

- **General learning formulation:**

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell_i(\theta^\top \Phi(x_i)) + \frac{\mu}{2} \|\theta\|_2^2 = \max_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^n \psi_i(\alpha_i) - \frac{1}{2\mu} \left\| \sum_{i=1}^n \alpha_i \Phi(x_i) \right\|_2^2$$

- **From primal to dual**

- $\ell_i$  smooth  $\Leftrightarrow \psi_i$  strongly convex
- $\ell_i$  strongly convex  $\Leftrightarrow \psi_i$  smooth

- **Applying coordinate descent in the dual**

- Nesterov (2012); Shalev-Shwartz and Zhang (2012)
- Linear convergence rate with simple iterations

# Dual stochastic coordinate ascent - IV

- **Dual formulation:**  $\max_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^n \psi_i(\alpha_i) - \frac{1}{2\mu} \left\| \sum_{i=1}^n \alpha_i \Phi(x_i) \right\|_2^2$
- **Stochastic coordinate descent:** at iteration  $t$ 
  - Choose a coordinate  $i$  at random
  - Optimize w.r.t.  $\alpha_i$ :  $\max_{\alpha_i \in \mathbb{R}} -\psi_i(\alpha_i) - \frac{1}{2\mu} \left\| \alpha_i \Phi(x_i) + \sum_{j \neq i} \alpha_j \Phi(x_j) \right\|_2^2$
  - Can be done by a **single access to  $\Phi(x_i)$**  and updating  $\sum_{j=1}^n \alpha_j \Phi(x_j)$
- **Convergence proof**
  - See Nesterov (2012); Shalev-Shwartz and Zhang (2012)
  - Similar linear convergence than SAG

# Randomized coordinate descent

## Proof - I

- **Simplest setting:** minimize  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  which is  $L$ -smooth
  - Local smoothness constants  $L_i = \sup_{\alpha \in \mathbb{R}^n} f''_{ii}(\alpha)$
  - $\max_{i \in \{1, \dots, n\}} L_i \leq L$  and  $L \leq \sum_{i=1}^n L_i$
  - NB: in dual problems in machine learning  $\max_{i \in \{1, \dots, n\}} L_i \propto R^2$
- **Algorithm:** at iteration  $t$ ,
  - Choose a coordinate  $i_t$  at random with probability  $p_i$
  - Local descent step:  $\alpha_t = \alpha_{t-1} - \frac{1}{L_{i_t}} f'(\alpha_{t-1})_{i_t} e_{i_t}$
- **Two choices for  $p_i$ :** (a) uniform or (b) proportional to  $L_i$



# Randomized coordinate descent

## Proof - II

- Iteration  $\alpha_t = \alpha_{t-1} - \frac{1}{L_{i_t}} f'(\alpha_{t-1})_{i_t} e_{i_t}$
- From smoothness,  $f(\alpha_t) \leq f(\alpha_{t-1}) - f'(\alpha_{t-1})^\top (\alpha_t - \alpha_{t-1}) + \frac{L_{i_t}}{2} \|\alpha_t - \alpha_{t-1}\|^2$   
leading to  $f(\alpha_t) \leq f(\alpha_{t-1}) - \frac{1}{2L_{i_t}} |f'(\alpha_{t-1})_{i_t}|^2$
- Taking expectations:  $\mathbb{E}[f(\alpha_t) | \mathcal{F}_{t-1}] \leq f(\alpha_{t-1}) - \sum_{i=1}^n \frac{p_i}{2L_i} |f'(\alpha_{t-1})_i|^2$

# Randomized coordinate descent

## Proof - II

- Iteration  $\alpha_t = \alpha_{t-1} - \frac{1}{L_{i_t}} f'(\alpha_{t-1})_{i_t} e_{i_t}$
- From smoothness,  $f(\alpha_t) \leq f(\alpha_{t-1}) - f'(\alpha_{t-1})^\top (\alpha_t - \alpha_{t-1}) + \frac{L_{i_t}}{2} \|\alpha_t - \alpha_{t-1}\|^2$   
leading to  $f(\alpha_t) \leq f(\alpha_{t-1}) - \frac{1}{2L_{i_t}} |f'(\alpha_{t-1})_{i_t}|^2$
- Taking expectations:  $\mathbb{E}[f(\alpha_t) | \mathcal{F}_{t-1}] \leq f(\alpha_{t-1}) - \sum_{i=1}^n \frac{p_i}{2L_i} |f'(\alpha_{t-1})_i|^2$
- If  $p_i = 1/n$  (uniform),  $\mathbb{E}[f(\alpha_t) | \mathcal{F}_{t-1}] \leq f(\alpha_{t-1}) - \frac{1}{2n \max_i L_i} \|f'(\alpha_{t-1})\|^2$   
With strong convexity:  $\mathbb{E}f(\alpha_t) \leq \mathbb{E}f(\alpha_{t-1}) - \frac{\mu}{n \max_i L_i} [\mathbb{E}f(\alpha_{t-1}) - f(\alpha^*)]$  leading  
to a linear convergence rate with factor  $1 - \frac{\mu}{n \max_i L_i}$

# Randomized coordinate descent

## Proof - II

- Iteration  $\alpha_t = \alpha_{t-1} - \frac{1}{L_{i_t}} f'(\alpha_{t-1})_{i_t} e_{i_t}$
- From smoothness,  $f(\alpha_t) \leq f(\alpha_{t-1}) - f'(\alpha_{t-1})^\top (\alpha_t - \alpha_{t-1}) + \frac{L_{i_t}}{2} \|\alpha_t - \alpha_{t-1}\|^2$   
leading to  $f(\alpha_t) \leq f(\alpha_{t-1}) - \frac{1}{2L_{i_t}} |f'(\alpha_{t-1})_{i_t}|^2$
- Taking expectations:  $\mathbb{E}[f(\alpha_t) | \mathcal{F}_{t-1}] \leq f(\alpha_{t-1}) - \sum_{i=1}^n \frac{p_i}{2L_i} |f'(\alpha_{t-1})_i|^2$
- If  $p_i = 1/n$  (uniform),  $\mathbb{E}[f(\alpha_t) | \mathcal{F}_{t-1}] \leq f(\alpha_{t-1}) - \frac{1}{2n \max_i L_i} \|f'(\alpha_{t-1})\|^2$   
With strong convexity:  $\mathbb{E}f(\alpha_t) \leq \mathbb{E}f(\alpha_{t-1}) - \frac{\mu}{n \max_i L_i} [\mathbb{E}f(\alpha_{t-1}) - f(\alpha^*)]$  leading  
to a linear convergence rate with factor  $1 - \frac{\mu}{n \max_i L_i}$
- If  $p_i = \frac{L_i}{\sum_{j=1}^n L_j}$ ,  $\mathbb{E}f(\alpha_t) \leq f(\alpha_{t-1}) - \frac{1}{2 \sum_{j=1}^n L_j} \|f'(\alpha_{t-1})\|^2$   
With strong convexity:  $\mathbb{E}f(\alpha_t) \leq \mathbb{E}f(\alpha_{t-1}) - \frac{\mu}{\sum_{j=1}^n L_j} [\mathbb{E}f(\alpha_{t-1}) - f(\alpha^*)]$  leading  
to a linear convergence rate with factor  $1 - \frac{\mu}{\sum_{j=1}^n L_j}$

# Randomized coordinate descent

## Discussion

- **Iteration**  $\alpha_t = \alpha_{t-1} - \frac{1}{L_{i_t}} f'(\alpha_{t-1})_{i_t} e_{i_t}$ 
  - If  $p_i = 1/n$  (uniform), linear rate  $1 - \frac{\mu}{n \max_i L_i}$
  - If  $p_i = \frac{L_i}{\sum_{j=1}^n L_j}$ , linear rate  $1 - \frac{\mu}{\sum_{j=1}^n L_j}$
- Best-case scenario:  $f''$  is diagonal, and  $L = \max_i L_i$
- Worst-case scenario:  $f''$  is constant and  $L = \sum_i L_i$

# Frank-Wolfe - conditional gradient - I

- **Goal:** minimize smooth convex function  $f(\theta)$  on compact set  $\mathcal{C}$
- **Standard result:** accelerated projected gradient descent with optimal rate  $O(1/t^2)$ 
  - Requires projection oracle:  $\arg \min_{\theta \in \mathcal{C}} \frac{1}{2} \|\theta - \eta\|^2$
- **Only availability of the linear oracle:**  $\arg \min_{\theta \in \mathcal{C}} \theta^\top \eta$ 
  - Many examples (sparsity, low-rank, large polytopes, etc.)
  - Iterative **Frank-Wolfe algorithm** (see, e.g., Jaggi, 2013, and references therein) *with geometric interpretation*

$$\begin{cases} \bar{\theta}_t \in \arg \min_{\theta \in \mathcal{C}} \theta^\top f'(\theta_{t-1}) \\ \theta_t = (1 - \rho_t)\theta_{t-1} + \rho_t \bar{\theta}_t \end{cases}$$

# Frank-Wolfe - conditional gradient - II

- **Convergence rates:**  $f(\theta_t) - f(\theta_*) \leq \frac{2L \text{diam}(\mathcal{C})^2}{t+1}$

$$\text{Iteration: } \begin{cases} \bar{\theta}_t \in \arg \min_{\theta \in \mathcal{C}} \theta^\top f'(\theta_{t-1}) \\ \theta_t = (1 - \rho_t)\theta_{t-1} + \rho_t \bar{\theta}_t \end{cases}$$

$$\text{From smoothness: } f(\theta_t) \leq f(\theta_{t-1}) + f'(\theta_{t-1})^\top [\rho_t(\bar{\theta}_t - \theta_{t-1})] + \frac{L}{2} \|\rho_t(\bar{\theta}_t - \theta_{t-1})\|^2$$

$$\text{From compactness: } f(\theta_t) \leq f(\theta_{t-1}) + f'(\theta_{t-1})^\top [\rho_t(\bar{\theta}_t - \theta_{t-1})] + \frac{L}{2} \rho_t^2 \text{diam}(\mathcal{C})^2$$

$$\text{From convexity, } f(\theta_t) - f(\theta_*) \leq f'(\theta_{t-1})^\top (\theta_{t-1} - \theta_*) \leq \max_{\theta \in \mathcal{C}} f'(\theta_{t-1})^\top (\theta_{t-1} - \theta),$$

which is equal to  $f'(\theta_{t-1})^\top (\theta_{t-1} - \bar{\theta}_t)$

$$\text{Thus, } f(\theta_t) \leq f(\theta_{t-1}) - \rho_t [f(\theta_{t-1}) - f(\theta_*)] + \frac{L}{2} \rho_t^2 \text{diam}(\mathcal{C})^2$$

$$\text{With } \rho_t = 2/(t+1): f(\theta_t) \leq \frac{2L \text{diam}(\mathcal{C})^2}{t+1} \text{ by direct expansion}$$

# Frank-Wolfe - conditional gradient - II

- **Convergence rates:**  $f(\theta_t) - f(\theta_*) \leq \frac{2L \text{diam}(\mathcal{C})^2}{t}$
- **Remarks and extensions**
  - Affine-invariant algorithms
  - Certified duality gaps and dual interpretations (Bach, 2015)
  - Adapted to very large polytopes
  - Line-search extensions: minimize quadratic upper-bound
  - Stochastic extensions (Lacoste-Julien et al., 2013)
  - Away and pairwise steps to avoid oscillations (Lacoste-Julien and Jaggi, 2015)

# Outline - I

## 1. Introduction

- Large-scale machine learning and optimization
- Classes of functions (convex, smooth, etc.)
- Traditional statistical analysis through Rademacher complexity

## 2. Classical methods for convex optimization

- Smooth optimization (gradient descent, Newton method)
- Non-smooth optimization (subgradient descent)
- Proximal methods

## 3. Non-smooth stochastic approximation

- Stochastic (sub)gradient and averaging
- Non-asymptotic results and lower bounds
- Strongly convex vs. non-strongly convex



# Outline - II

## 4. Classical stochastic approximation

- Asymptotic analysis
- Robbins-Monro algorithm
- Polyak-Rupert averaging

## 5. Smooth stochastic approximation algorithms

- Non-asymptotic analysis for smooth functions
- Logistic regression
- Least-squares regression without decaying step-sizes

## 6. Finite data sets

- Gradient methods with exponential convergence rates
- Convex duality
- (Dual) stochastic coordinate descent - Frank-Wolfe

# Subgradient descent for machine learning

- **Assumptions** ( $f$  is the expected risk,  $\hat{f}$  the empirical risk)
  - “Linear” predictors:  $\theta(x) = \theta^\top \Phi(x)$ , with  $\|\Phi(x)\|_2 \leq R$  a.s.
  - $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi(x_i)^\top \theta)$
  - $G$ -Lipschitz loss:  $f$  and  $\hat{f}$  are  $GR$ -Lipschitz on  $\Theta = \{\|\theta\|_2 \leq D\}$

- **Statistics:** with probability greater than  $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{GRD}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- **Optimization:** after  $t$  iterations of subgradient method

$$\hat{f}(\hat{\theta}) - \min_{\eta \in \Theta} \hat{f}(\eta) \leq \frac{GRD}{\sqrt{t}}$$

- $t = n$  iterations, with total running-time complexity of  $O(n^2d)$

# Stochastic subgradient “descent” / method

- **Assumptions**

- $f_n$  convex and  $B$ -Lipschitz-continuous on  $\{\|\theta\|_2 \leq D\}$
- $(f_n)$  i.i.d. functions such that  $\mathbb{E}f_n = f$
- $\theta_*$  global optimum of  $f$  on  $\{\|\theta\|_2 \leq D\}$

- **Algorithm:**  $\theta_n = \Pi_D \left( \theta_{n-1} - \frac{2D}{B\sqrt{n}} f'_n(\theta_{n-1}) \right)$

- **Bound:**

$$\mathbb{E}f \left( \frac{1}{n} \sum_{k=0}^{n-1} \theta_k \right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$$

- “Same” three-line proof as in the deterministic case
- **Minimax rate** (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
- Running-time complexity:  $O(dn)$  after  $n$  iterations

# Summary of new results (Bach and Moulines, 2011)

- Stochastic gradient descent with learning rate  $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old:  $O(n^{-1})$  rate achieved **without** averaging for  $\alpha = 1$
  - New:  $O(n^{-1})$  rate achieved **with** averaging for  $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants
  - Forgetting of initial conditions
  - Robustness to the choice of  $C$
- **Convergence rates** for  $\mathbb{E}\|\theta_n - \theta_*\|^2$  and  $\mathbb{E}\|\bar{\theta}_n - \theta_*\|^2$ 
  - no averaging:  $O\left(\frac{\sigma^2 \gamma_n}{\mu}\right) + O(e^{-\mu n \gamma_n})\|\theta_0 - \theta_*\|^2$
  - averaging:  $\frac{\text{tr } H(\theta_*)^{-1}}{n} + \mu^{-1}O(n^{-2\alpha} + n^{-2+\alpha}) + O\left(\frac{\|\theta_0 - \theta_*\|^2}{\mu^2 n^2}\right)$

# Least-mean-square algorithm

- **Least-squares:**  $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \Phi(x_n), \theta \rangle)^2]$  with  $\theta \in \mathbb{R}^d$ 
  - SGD = least-mean-square algorithm (see, e.g., Macchi, 1995)
  - usually studied without averaging and decreasing step-sizes
  - with strong convexity assumption  $\mathbb{E}[\Phi(x_n) \otimes \Phi(x_n)] = H \succcurlyeq \mu \cdot \text{Id}$
- **New analysis for averaging and constant step-size**  $\gamma = 1/(4R^2)$ 
  - Assume  $\|\Phi(x_n)\| \leq R$  and  $|y_n - \langle \Phi(x_n), \theta_* \rangle| \leq \sigma$  almost surely
  - **No assumption regarding lowest eigenvalues of  $H$**
  - Main result: 
$$\mathbb{E}f(\bar{\theta}_{n-1}) - f(\theta_*) \leq \frac{4\sigma^2 d}{n} + \frac{4R^2 \|\theta_0 - \theta_*\|^2}{n}$$
- **Matches statistical lower bound** (Tsybakov, 2003)
  - Non-asymptotic robust version of Györfi and Walk (1996)

# Choice of support point for online Newton step

- **Two-stage procedure**

- (1) Run  $n/2$  iterations of averaged SGD to obtain  $\tilde{\theta}$

- (2) Run  $n/2$  iterations of averaged constant step-size LMS

- Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)

- **Provable convergence rate of  $O(d/n)$**  for logistic regression

- Additional assumptions but no **strong convexity**

- **Update at each iteration using the current averaged iterate**

- Recursion: 
$$\theta_n = \theta_{n-1} - \gamma [f'_n(\bar{\theta}_{n-1}) + f''_n(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1})]$$

- No provable convergence rate (yet) but best practical behavior

- Note (dis)similarity with regular SGD:  $\theta_n = \theta_{n-1} - \gamma f'_n(\theta_{n-1})$

# Stochastic average gradient (Le Roux, Schmidt, and Bach, 2012)

- **Stochastic average gradient (SAG) iteration**
  - Keep in memory the gradients of all functions  $f_i, i = 1, \dots, n$
  - Random selection  $i(t) \in \{1, \dots, n\}$  with replacement
  - Iteration:  $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^n y_i^t$  with  $y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$
- Stochastic version of incremental average gradient (Blatt et al., 2008)
- Extra memory requirement
  - Supervised machine learning
    - If  $f_i(\theta) = \ell_i(y_i, \Phi(x_i)^\top \theta)$ , then  $f'_i(\theta) = \ell'_i(y_i, \Phi(x_i)^\top \theta) \Phi(x_i)$
    - Only need to store  $n$  real numbers

# Summary of rates of convergence

- Problem parameters
  - $D$  diameter of the domain
  - $B$  Lipschitz-constant
  - $L$  smoothness constant
  - $\mu$  strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$ stochastic: $BD/\sqrt{n}$	deterministic: $B^2/(t\mu)$ stochastic: $B^2/(n\mu)$
smooth	deterministic: $LD^2/t^2$ stochastic: $LD^2/\sqrt{n}$ finite sum: $n/t$	deterministic: $\exp(-t\sqrt{\mu/L})$ stochastic: $L/(n\mu)$ finite sum: $\exp(-\min\{1/n, \mu/L\}t)$
quadratic	deterministic: $LD^2/t^2$ stochastic: $d/n + LD^2/n$	deterministic: $\exp(-t\sqrt{\mu/L})$ stochastic: $d/n + LD^2/n$



# Conclusions

## Machine learning and convex optimization

- **Statistics with or without optimization?**
  - **Significance** of mixing algorithms with analysis
  - **Benefits** of mixing algorithms with analysis
- **Open problems**
  - Non-parametric stochastic approximation
  - Characterization of implicit regularization of online methods
  - Structured prediction
  - Going beyond a single pass over the data (testing performance)
  - Further links between convex optimization and online learning/bandits
  - Parallel and distributed optimization
  - Non-convex optimization

# References

- A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.
- Alekh Agarwal and Leon Bottou. A lower bound for the optimization of finite sums. *arXiv preprint arXiv:1410.0723*, 2014.
- R. Aguech, E. Moulines, and P. Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control and Optimization*, 39(3):872–899, 2000.
- Y. Arjevani and O. Shamir. Dimension-free iteration complexity of finite sum optimization problems. In *Advances In Neural Information Processing Systems*, 2016.
- R. Babanezhad, M. O. Ahmed, A. Virani, M. W. Schmidt, J. Konečný, and S. Sallinen. Stopwasting my gradients: Practical SVRG. In *Advances in Neural Information Processing Systems*, 2015.
- F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. ISSN 1935-7524.
- F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. Technical Report 00804431, HAL, 2013.
- F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*, 2011.
- F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate  $o(1/n)$ . Technical Report 00831977, HAL, 2013.

- F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization, 2012a.
- Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.
- Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012b.
- P. Balamurugan and F. Bach. Stochastic variance reduction methods for saddle-point problems. Technical Report 01319293, HAL, 2016.
- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer Publishing Company, Incorporated, 2012.
- D. P. Bertsekas. *Nonlinear programming*. Athena scientific, 1999.
- D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2008.
- V. S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.
- L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Adv. NIPS*, 2008.
- L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.
- S. Boucheron and P. Massart. A high-dimensional wilks phenomenon. *Probability theory and related*

*fields*, 150(3-4):405–433, 2011.

- S. Boucheron, O. Bousquet, G. Lugosi, et al. Theory of classification: A survey of some recent advances. *ESAIM Probability and statistics*, 9:323–375, 2005.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003.
- S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.
- N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9):2050–2057, 2004.
- P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.
- A. d’Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183, 2008.
- A. Defazio, J. Domke, and T. S. Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *Proc. ICML*, 2014a.
- Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014b.
- A. Défossez and F. Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. 2015.
- A. Dieuleveut and F. Bach. Non-parametric Stochastic Approximation with Large Step sizes. Technical report, ArXiv, 2014.

- A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. Technical Report 1602.05419, arXiv, 2016.
- J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009. ISSN 1532-4435.
- M. Duflo. *Algorithmes stochastiques*. Springer-Verlag, 1996.
- V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.
- N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. *arXiv preprint arXiv:1504.01577*, 2015.
- R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. In *Proceedings of the Conference on Learning Theory*, 2015.
- S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization. *Optimization Online*, July, 2010.
- Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.
- M. Gurbuzbalaban, A. Ozdaglar, and P. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. Technical Report 1506.02081, arXiv, 2015.
- L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.
- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization.

*Machine Learning*, 69(2):169–192, 2007.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

Chonghai Hu, James T Kwok, and Weike Pan. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, volume 22, pages 781–789, 2009.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 427–435, 2013.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Takafumi Kanamori and Hidetoshi Shimodaira. Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of statistical planning and inference*, 116(1):149–162, 2003.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, second edition, 2003.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an  $o(1/t)$  convergence rate for projected stochastic subgradient descent. Technical Report 1212.2002, ArXiv, 2012.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate {Frank-Wolfe} optimization for structural {SVMs}. In *Proceedings of The 30th International Conference on Machine Learning*, pages 53–61, 2013.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A):

365–397, 2012.

- G. Lan. An optimal randomized incremental gradient method. Technical Report 1507.02000, arXiv, 2015.
- Guanghui Lan, Arkadi Nemirovski, and Alexander Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.
- N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Adv. NIPS*, 2012.
- N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. Technical Report 00674995, HAL, 2013.
- R. Leblond, F. Pedregosa, and S. Lacoste-Julien. Asaga: Asynchronous parallel Saga. Technical Report 1606.04809, arXiv, 2016.
- H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- O. Macchi. *Adaptive processing: The least mean squares approach with applications in transmission*. Wiley West Sussex, 1995.
- J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.
- P. Massart. *Concentration Inequalities and Model Selection: Ecole d’été de Probabilités de Saint-Flour 23*. Springer, 2003.
- R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine*

*Learning Research*, 4:839–860, 2003.

- A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. *Stochastic Optimization: Algorithms and Applications*, pages 263–304, 2000.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley & Sons, 1983.
- Y. Nesterov. A method for solving a convex programming problem with rate of convergence  $O(1/k^2)$ . *Soviet Math. Doklady*, 269(3):543–547, 1983.
- Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.
- Y. Nesterov. Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep, 76*, 2007.
- Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.
- Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. SIAM studies in Applied Mathematics, 1994.
- Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6): 1559–1568, 2008. ISSN 0005-1098.
- Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.



- A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.
- I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):pp. 1679–1706, 1994. URL <http://www.jstor.org/stable/2244912>.
- B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *Information Theory, IEEE Transactions on*, 57(10):7036–7056, 2011.
- H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951a.
- H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951b.
- Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*, pages 111–135. Springer, 1985.
- R Tyrrell Rockafellar. *Convex Analysis*. Number 28. Princeton University Press, 1997.
- D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.
- M. Schmidt, N. Le Roux, and F. Bach. Optimization with approximate gradients. Technical report, HAL, 2011.

- B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.
- S. Shalev-Shwartz. Sdca without duality, regularization, and individual convexity. Technical Report 1602.01582, arXiv, 2016.
- S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proc. ICML*, 2008.
- S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. Technical Report 1209.1873, Arxiv, 2012.
- S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proc. ICML*, 2014.
- S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. ICML*, 2007.
- S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *proc. COLT*, 2009.
- O. Shamir. Without-replacement sampling for stochastic gradient methods: Convergence results and application to distributed optimization. Technical Report 1603.00570, arXiv, 2016.
- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- Naum Zuselevich Shor, Krzysztof C. Kiwiel, and Andrzej Ruszcay?ski. *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc., 1985.
- M.V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

- K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. 2008.
- P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.
- I. Tsochantaridis, Thomas Joachims, T., Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.
- A. B. Tsybakov. Optimal rates of aggregation. In *Proc. COLT*, 2003.
- A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge Univ. press, 2000.
- L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010. ISSN 1532-4435.
- L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.
- L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, 2013.