Statistical machine learning and convex optimization

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Université Paris-Sud (Mathematics dpt.) - Spring 2016

Slides available: `www.di.ens.fr/~fbach/cours_orsay_2016_slides.pdf`

Statistical machine learning and convex optimization

- **Six classes**

  1. Thursday February 18, 2pm to 5pm
  2. Thursday February 25, 2pm to 5pm
  3. Thursday March 17, 2pm to 5pm
  4. Thursday March 24, 2pm to 5pm
  5. Thursday April 14, 2pm to 5pm
  6. Monday April 18, 2pm to 5pm

- **Evaluations**

  1. Poster session presenting research papers
     – Thursday April 21, 2pm-5pm
     – See course web page for list of papers and instructions
  2. Scribe notes for a lecture

Statistical machine learning and convex optimization

Poster session - April, 21 - 2pm to 5pm

- Prepare 8-12 slides to present
  - Main ideas of the paper
  - Its relationship to the class
  - Potentially the main elements of proofs
  - If applicable a simple simulation

- Slides may be prepared in French or English

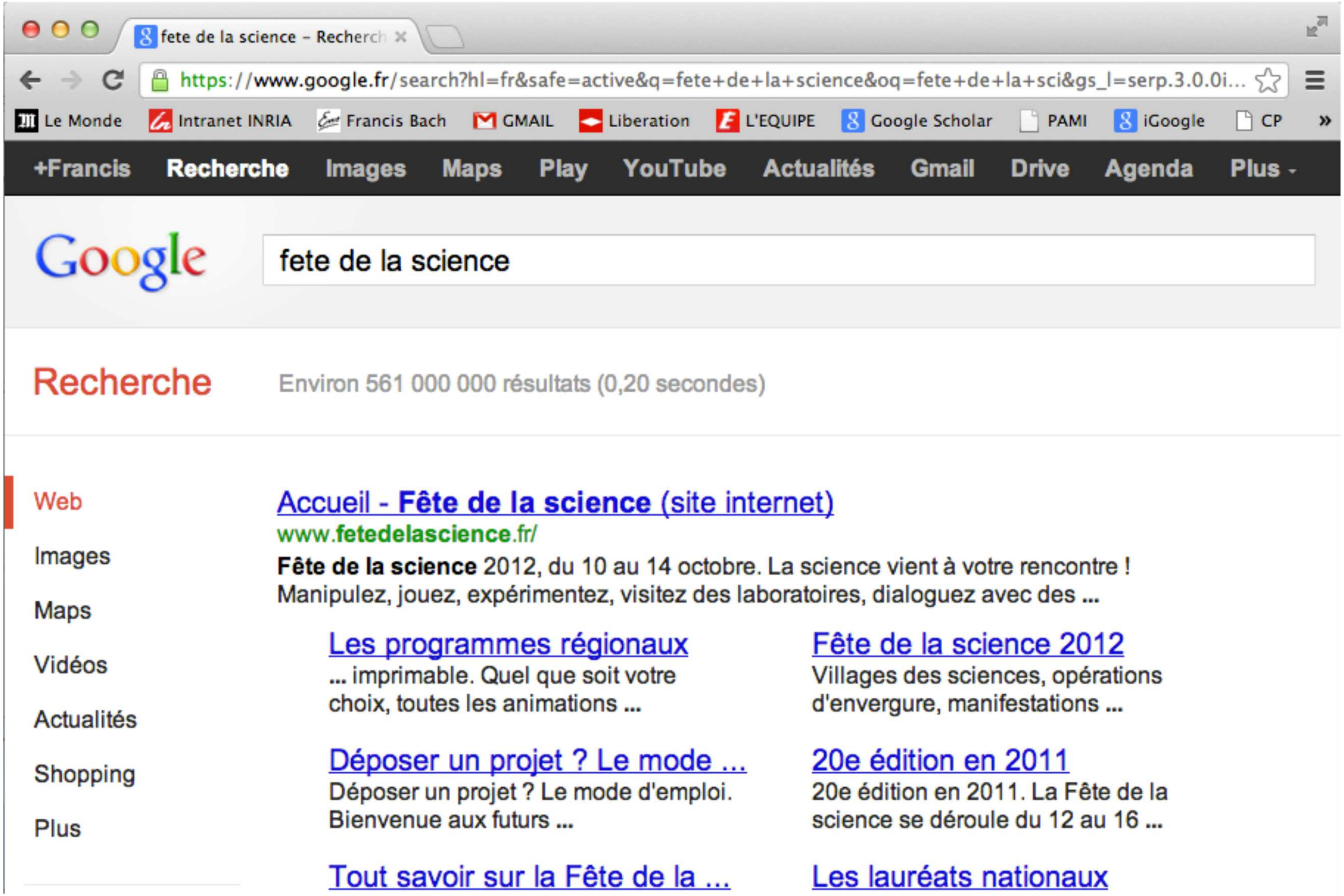
- In some cases it may be worth selecting a relevant subset of results to present.

- Two target audiences: (a) lecturer and (b) other students
  - Prepare a 10 minute walk-through

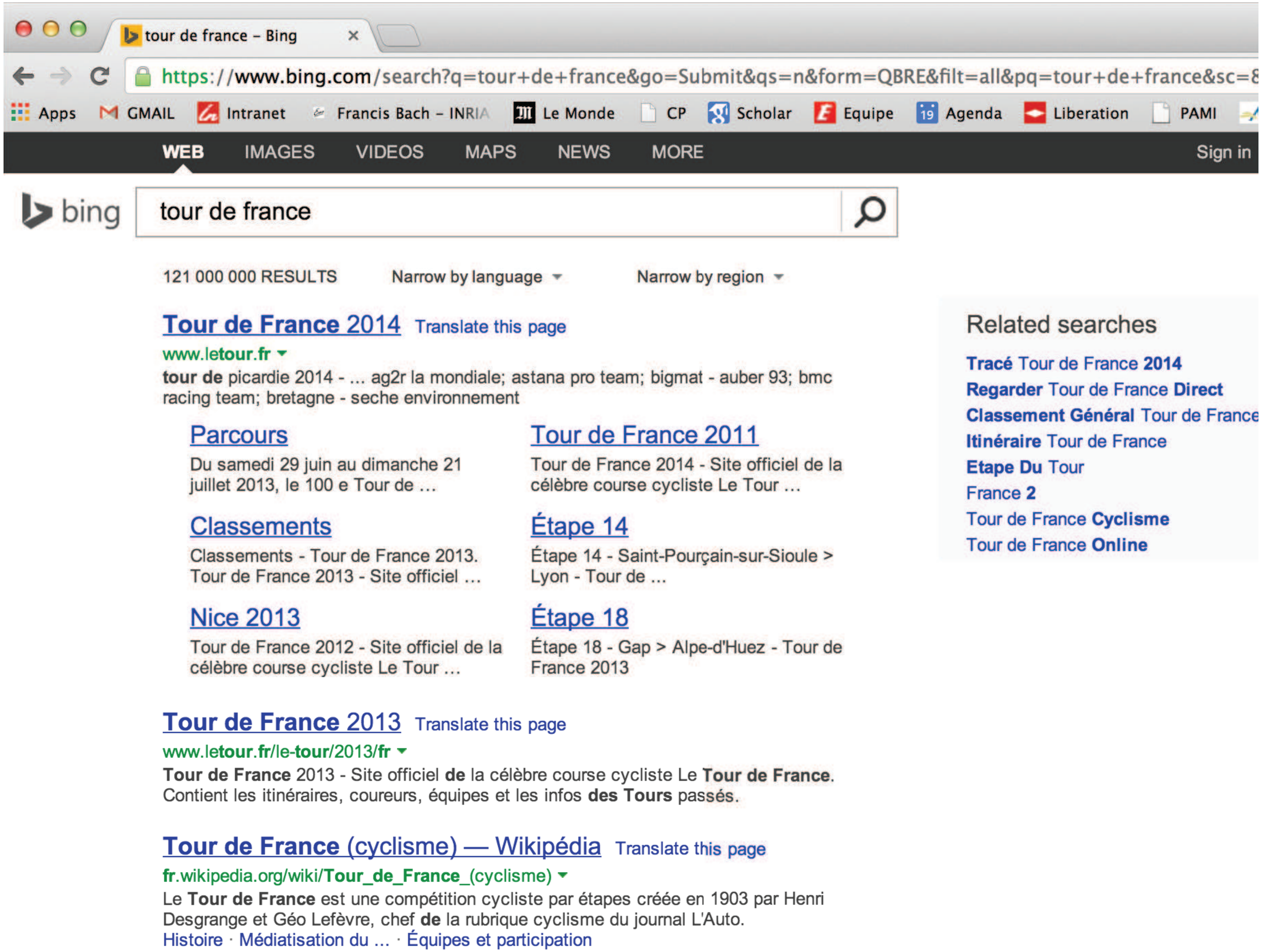
"Big data" revolution?
A new scientific context

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

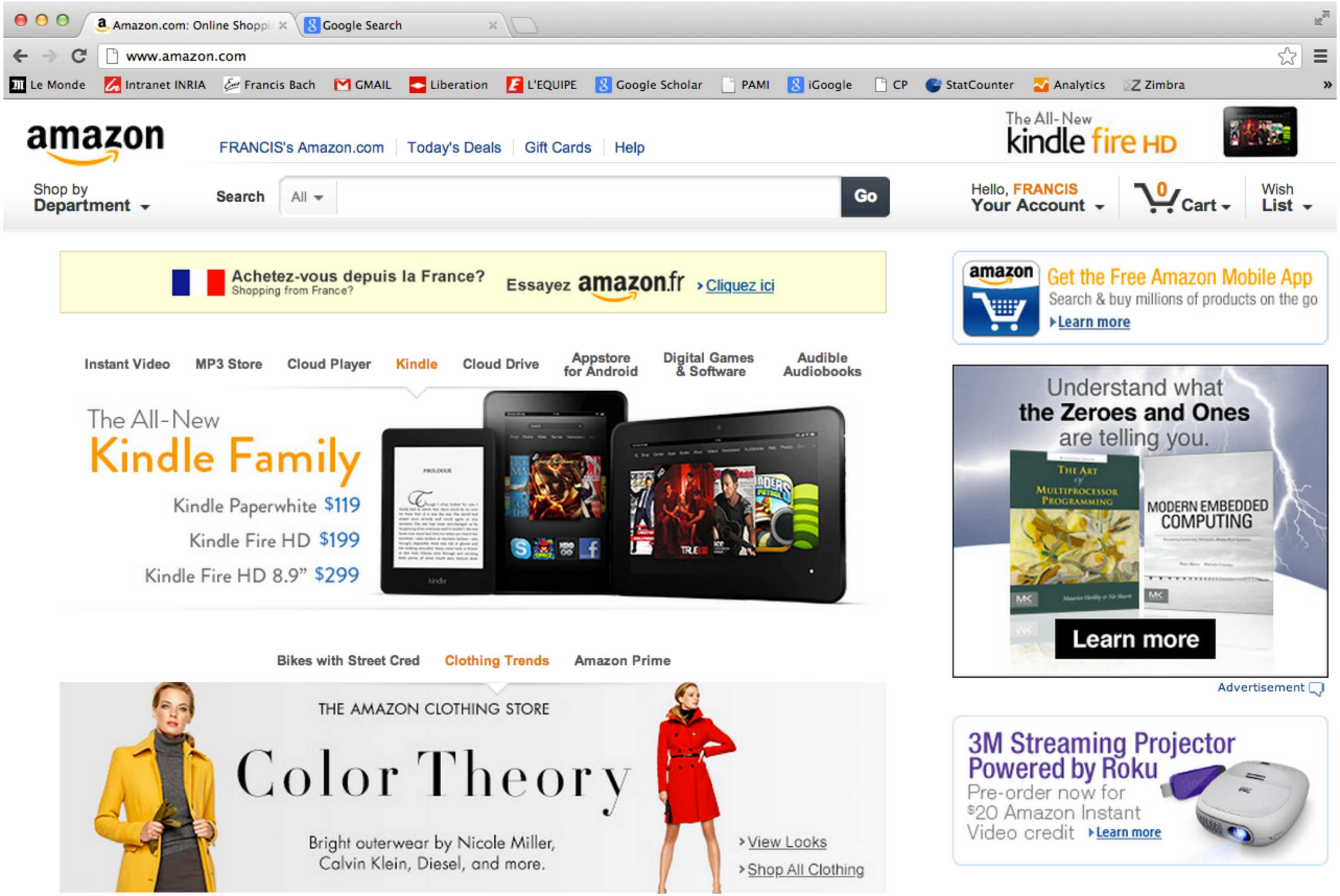
Search engines - Advertising



Search engines - Advertising

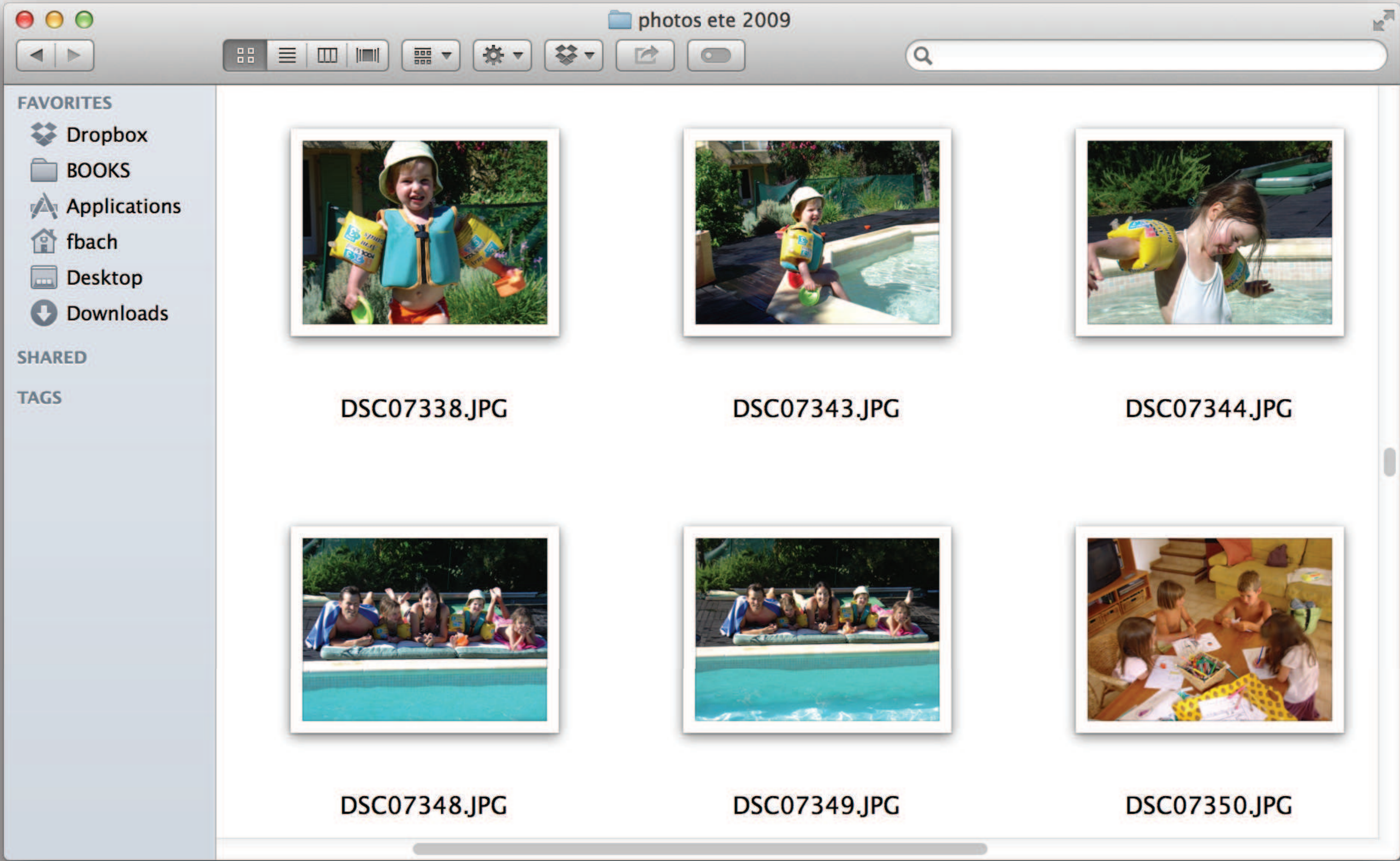


Marketing - Personalized recommendation

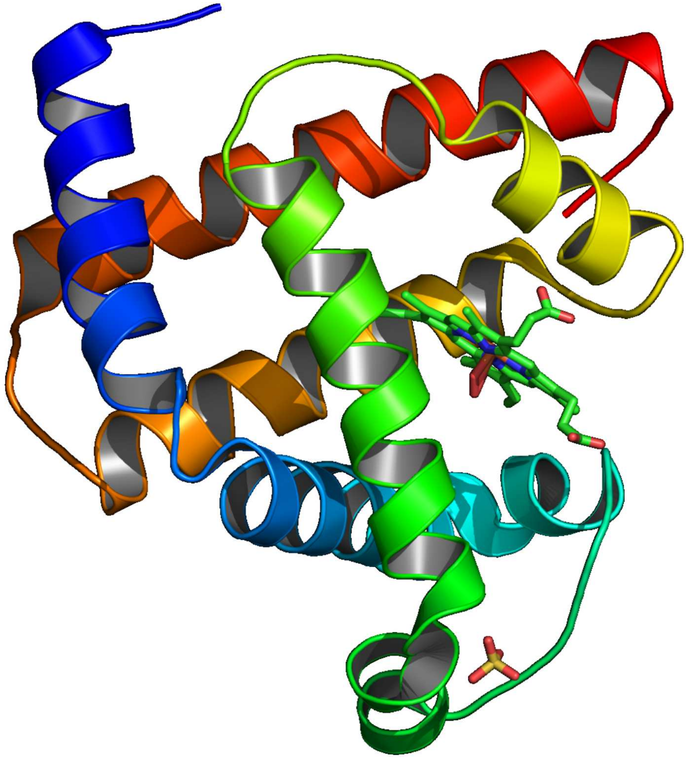


Visual object recognition

Personal photos



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, 1951)
  - Mixing statistics and optimization

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. **Classical stochastic approximation**

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

Outline - II

4. **Non-smooth stochastic approximation**

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

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

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

convex data fitting term $+$ regularizer

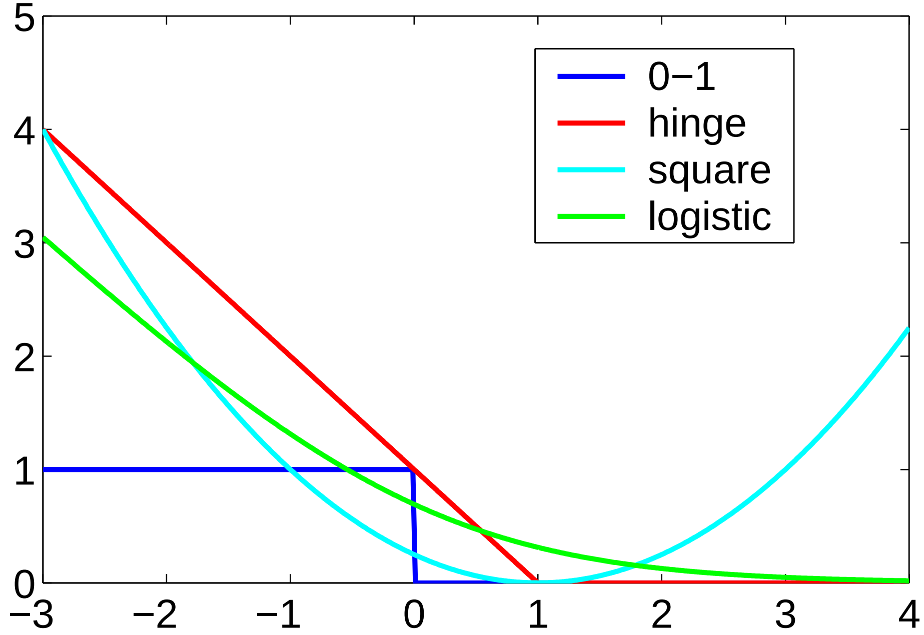
Usual losses

- **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: $\ell(y\,\theta^\top \Phi(x)) = 1_{y\,\theta^\top \Phi(x) < 0}$
  - 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: ℓ_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

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

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

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

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

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

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

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

convex data fitting term + constraint

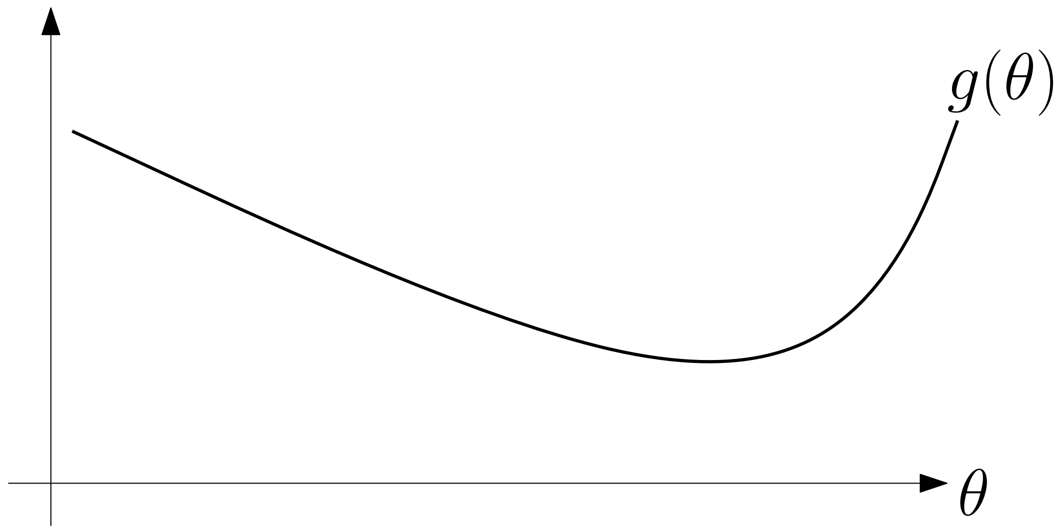
- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$ training cost
- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$ testing cost
- **Two fundamental questions**: (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$
  - **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 $\Phi(x) \in \mathbb{R}^d$: $\|\Phi(x)\|_2 \leqslant R$
- 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

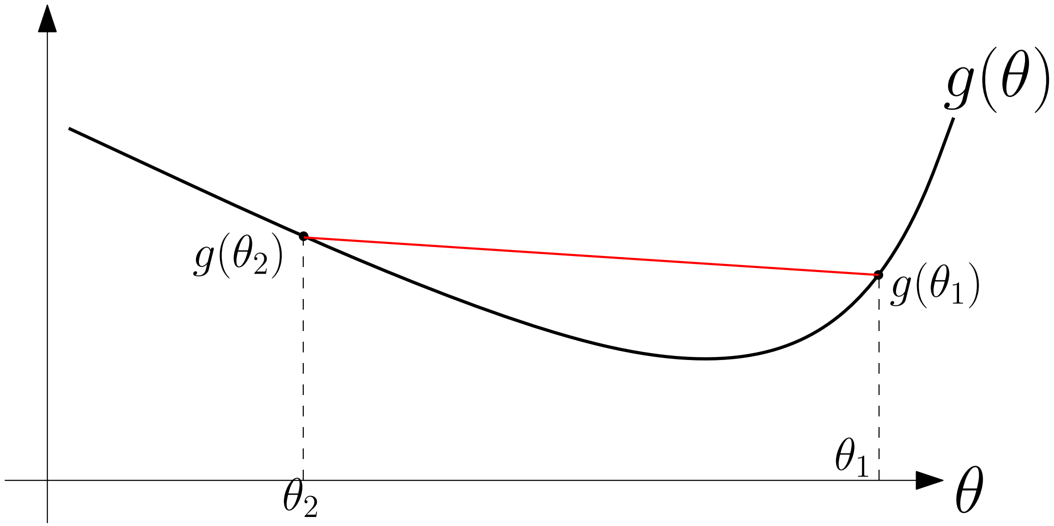
Convexity

- **Global definitions**



Convexity

- **Global definitions (full domain)**

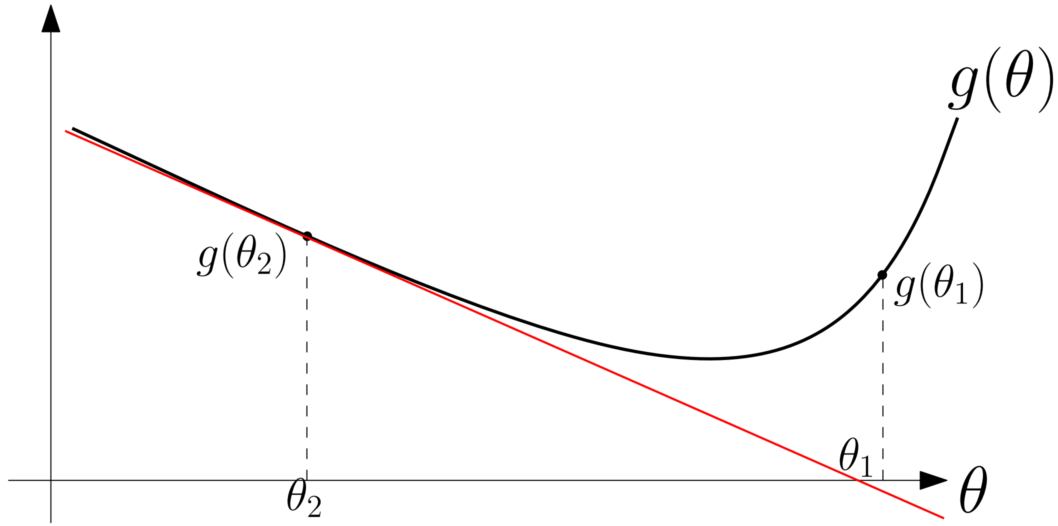


  – Not assuming differentiability:

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

Convexity

- **Global definitions (full domain)**



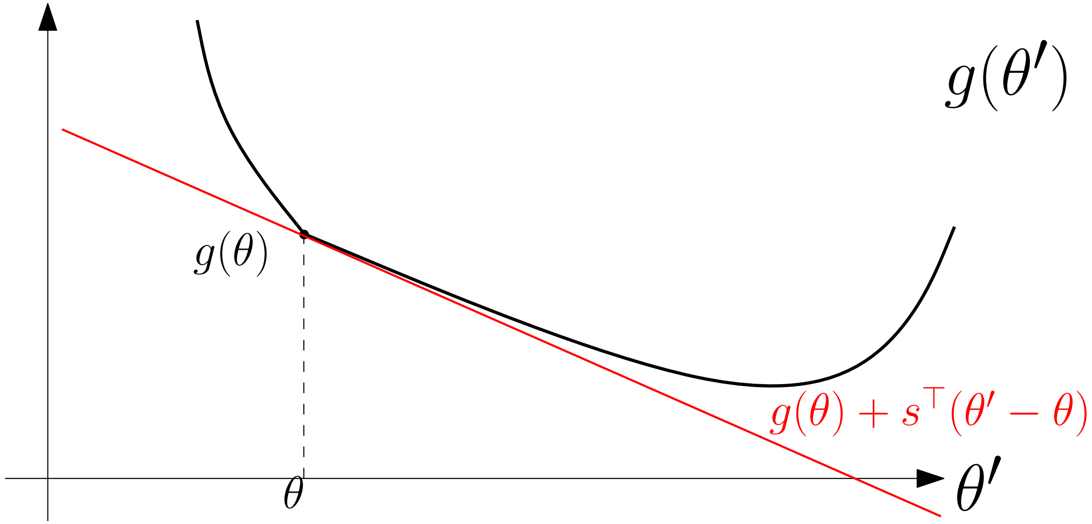
  - Assuming differentiability:

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

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

Subgradients and subdifferentials

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



  – $s \in \mathbb{R}^d$ is a subgradient of g at θ if and only if

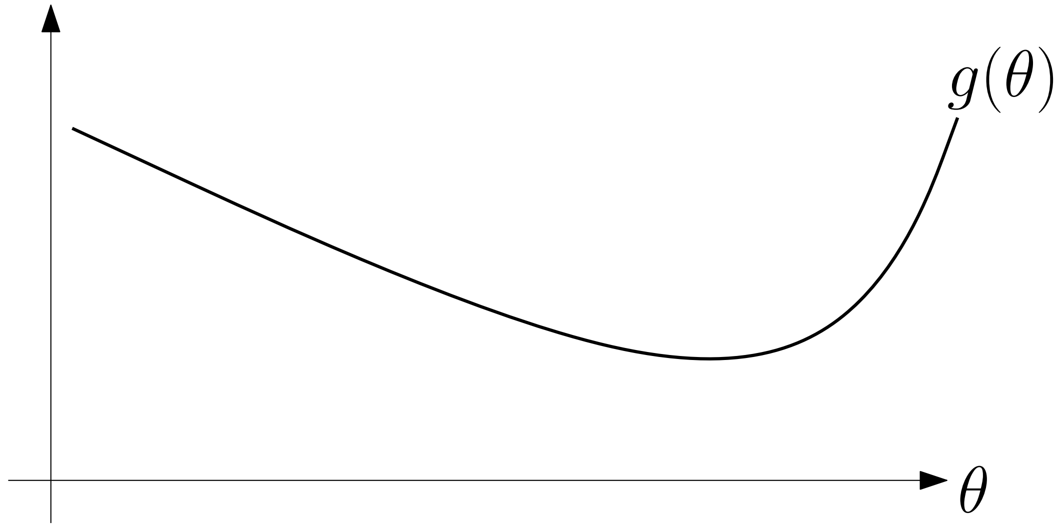
$$\forall \theta' \in \mathbb{R}^d, g(\theta') \geqslant g(\theta) + s^\top(\theta' - \theta)$$

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

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

Convexity

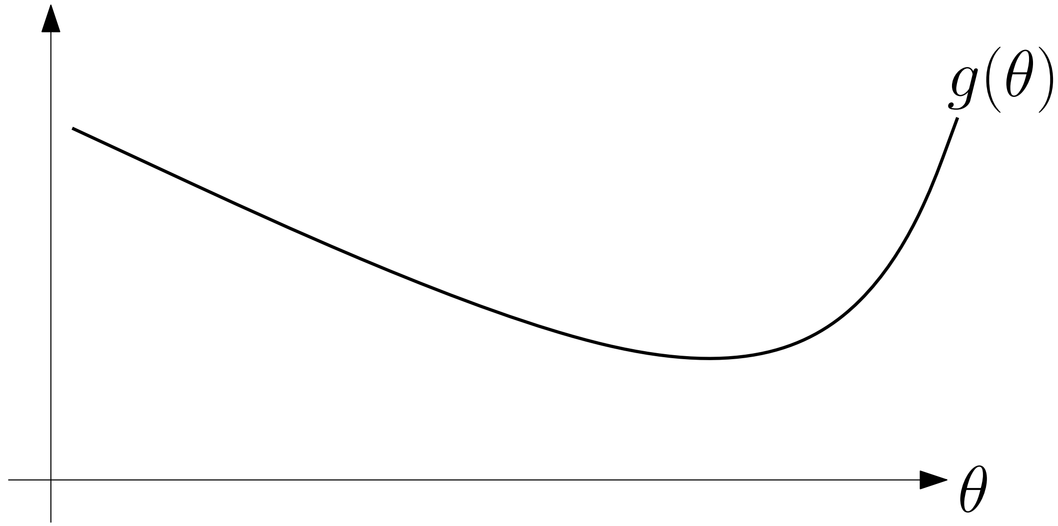
- **Global definitions (full domain)**



- **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 if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

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

$$\Leftrightarrow$$

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

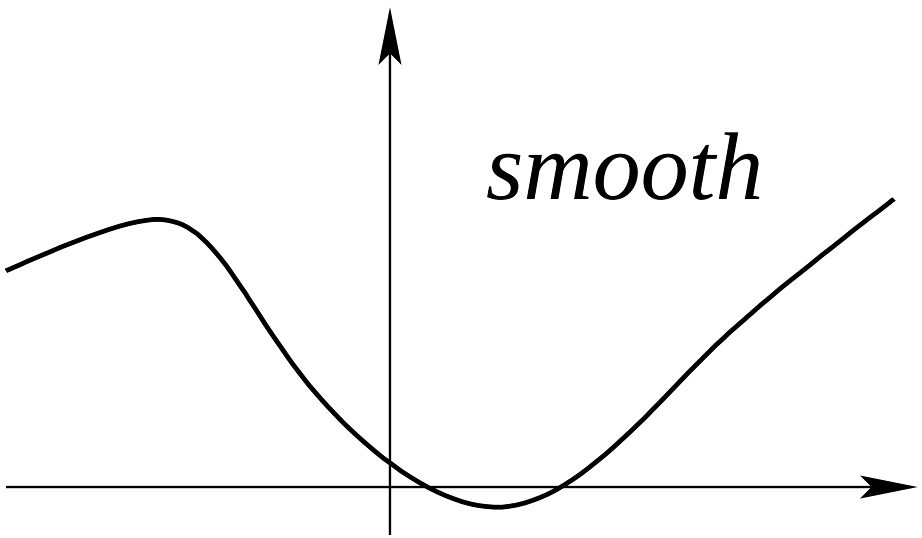
- **Machine learning**
  - with $g(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
  - G-Lipschitz loss and R-bounded data: $B = GR$

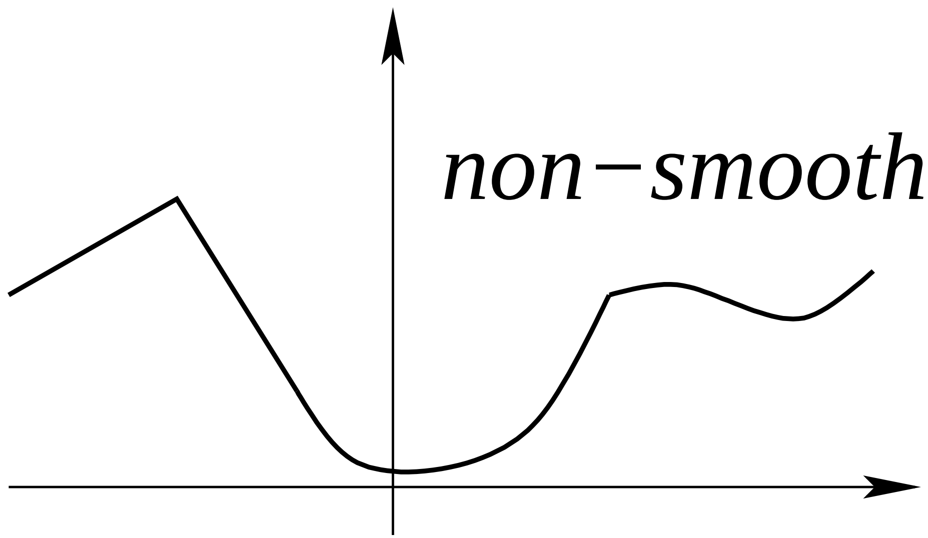
Smoothness and strong convexity

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

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

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





Smoothness and strong convexity

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

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

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

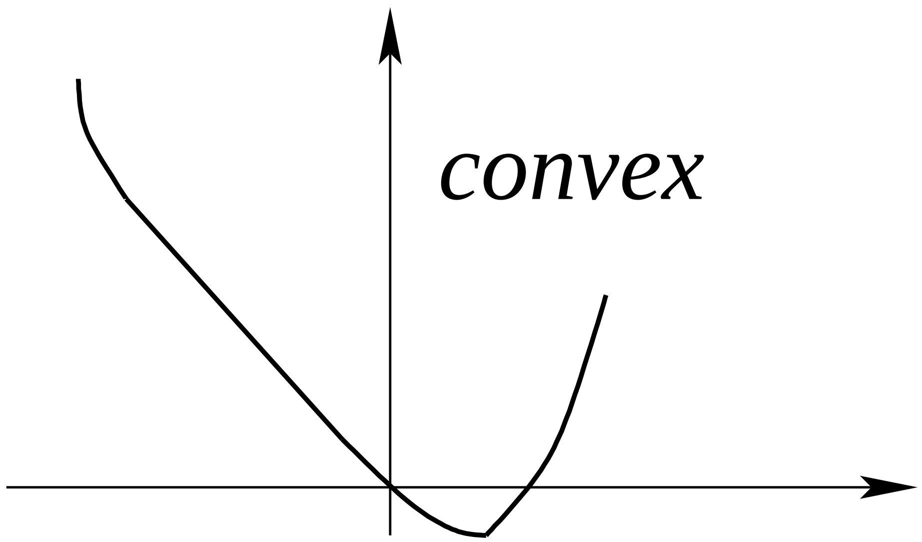
- **Machine learning**
  - with $g(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$
  - Hessian $\approx$ covariance matrix $\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top$
  - L_{loss}-smooth loss and R-bounded data: $L = L_{\text{loss}} R^2$

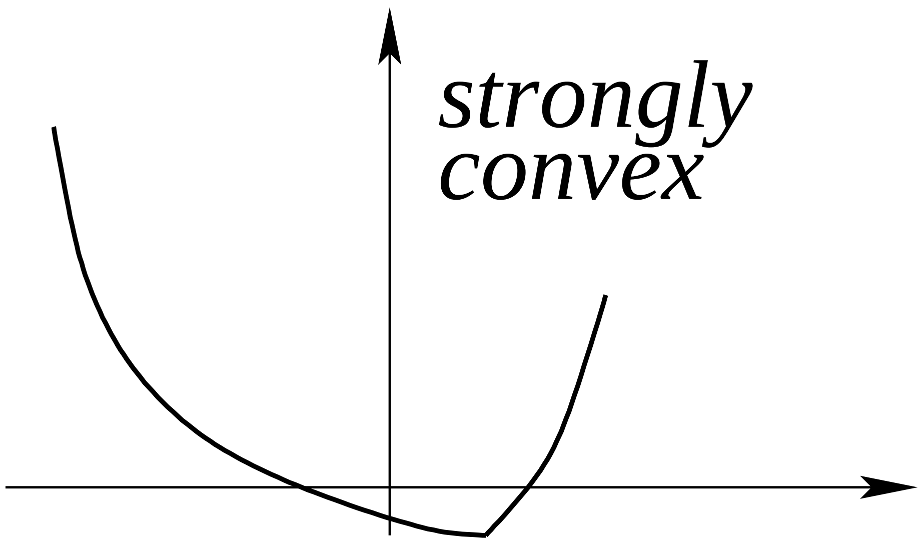
Smoothness and strong convexity

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

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

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



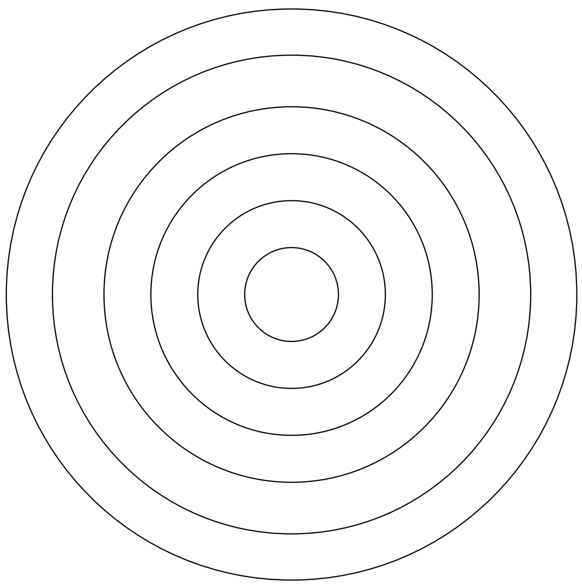


Smoothness and strong convexity

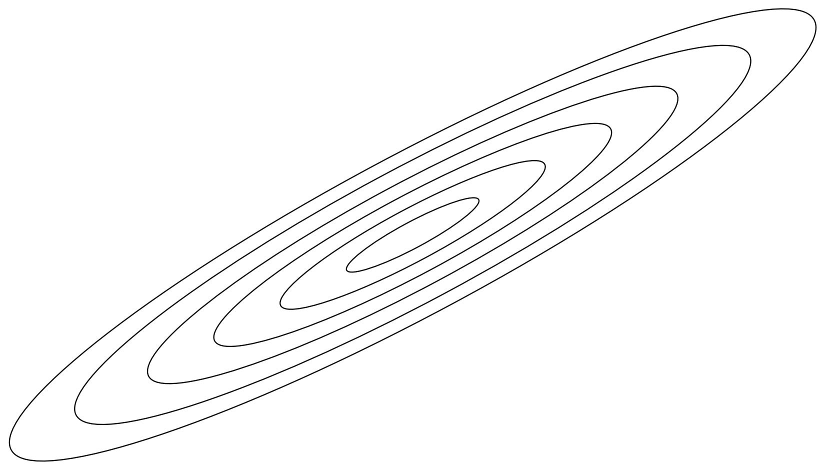
- A function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if

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

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



(large μ)



(small μ)

Smoothness and strong convexity

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

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

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

- **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 $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if

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

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

- **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 μ is small

Summary of smoothness/convexity assumptions

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

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

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

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

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

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

Analysis of empirical risk minimization

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

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \left[f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right] + \left[\min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]$$

Estimation error 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**: $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leqslant D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \Big[f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \Big] + \Big[\min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \Big]$$

Estimation error — Approximation error

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

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

Analysis of empirical risk minimization

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

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \left[f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \right] + \left[\min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \right]$$

Estimation error Approximation error

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

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

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

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

Analysis of empirical risk minimization

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

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} f(\theta) = \Big[f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \Big] + \Big[\min_{\theta \in \Theta} f(\theta) - \min_{\theta \in \mathbb{R}^d} f(\theta) \Big]$$

Estimation error   Approximation error

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

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

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

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

$$
\begin{aligned}
f(\theta) - \hat{f}(\theta) &= \frac{1}{2}\theta^\top \bigg(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top - \mathbb{E}\Phi(X)\Phi(X)^\top \bigg)\theta \\
&\quad - \theta^\top \bigg(\frac{1}{n}\sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E} Y\Phi(X) \bigg) + \frac{1}{2}\bigg(\frac{1}{n}\sum_{i=1}^n y_i^2 - \mathbb{E}Y^2 \bigg), \\
\sup_{\|\theta\|_2 \leqslant D} |f(\theta) - \hat{f}(\theta)| &\leqslant \frac{D^2}{2} \bigg\| \frac{1}{n}\sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top - \mathbb{E}\Phi(X)\Phi(X)^\top \bigg\|_{\mathrm{op}} \\
&\quad + D \bigg\| \frac{1}{n}\sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E} Y\Phi(X) \bigg\|_2 + \frac{1}{2}\bigg| \frac{1}{n}\sum_{i=1}^n y_i^2 - \mathbb{E}Y^2 \bigg|, \\
\sup_{\|\theta\|_2 \leqslant D} |f(\theta) - \hat{f}(\theta)| &\leqslant O(1/\sqrt{n}) \text{ with high probability from 3 concentrations}
\end{aligned}
$$

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 $\|\Phi(x)\|_2 \leqslant R$ a.s.
  - G-Lipschitz loss: f and $\hat{f}$ are GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leqslant D\}$
  - 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 $\|\Phi(x)\|_2 \leqslant R$ a.s.
  - G-Lipschitz loss: f and $\hat{f}$ are GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leqslant D\}$
  - No assumptions regarding convexity

- With probability greater than $1 - \delta$

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

- Expectated estimation error: $\mathbb{E}\big[\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)|\big] \leqslant \dfrac{4\ell_0 + 4GRD}{\sqrt{n}}$

- 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 $f'_i(\theta)$

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

Rademacher complexity

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

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

  – 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 ε

  – "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**:

$$\mathbb{E}\big[\sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) \big] = \mathbb{E}\big[\sup_{\theta \in \Theta} \hat{f}(\theta) - f(\theta) \big] \leqslant 2 R_n$$

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

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

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

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

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

Bounding the Rademacher average - I

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

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

- Using Ledoux-Talagrand concentration results for Rademacher averages (since φ_i is G-Lipschitz), we get (Meir and Zhang, 2003):

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

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

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

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

- **Proof by induction on n**
  - $n = 0$: trivial

- From n to $n + 1$: see next slide

From n to $n+1$

$$\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^{n+1}\varepsilon_i\varphi_i(a_i(\theta))\right]$$

$$= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\left[\sup_{\theta,\theta'\in\Theta}\frac{b(\theta)+b(\theta')}{2} + \sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta))-\varphi_{n+1}(a_{n+1}(\theta'))}{2}\right]$$

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

$$\leqslant \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\left[\sup_{\theta,\theta'\in\Theta}\frac{b(\theta)+b(\theta')}{2} + \sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2} + \frac{|a_{n+1}(\theta)-a_{n+1}(\theta')|}{2}\right]$$

$$= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\mathbb{E}_{\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta} b(\theta) + \varepsilon_{n+1}a_{n+1}(\theta) + \sum_{i=1}^{n}\varepsilon_i\varphi_i(a_i(\theta))\right]$$

$$\leqslant \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta} b(\theta) + \varepsilon_{n+1}a_{n+1}(\theta) + \sum_{i=1}^{n}\varepsilon_i a_i(\theta)\right] \text{ by recursion}$$

Bounding the Rademacher average - II

- We have:

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

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

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

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

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

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

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

- **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

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

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

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

- **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 $\|\Phi(x)\|_2 \leqslant R$ a.s.
  - G-Lipschitz loss: f and $\hat{f}$ are GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leqslant D\}$
  - No assumptions regarding convexity

- With probability greater than $1 - \delta$

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

- Expectated estimation error: $\mathbb{E}\big[\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)|\big] \leqslant \dfrac{4(\ell_0 + GRD)}{\sqrt{n}}$

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

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

- More refined/direct bound:

$$\begin{aligned} f(\hat{\theta}) - f(\mathbb{E}z) &= \frac{1}{2}(\hat{\theta} - \mathbb{E}z)^2 \\ \mathbb{E}\big[f(\hat{\theta}) - f(\mathbb{E}z)\big] &= \frac{1}{2}\mathbb{E}\bigg(\frac{1}{n}\sum_{i=1}^n z_i - \mathbb{E}z\bigg)^2 = \frac{1}{2n}\operatorname{var}(z) \end{aligned}$$

- 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 $\theta \in \mathbb{R}^d$,

$$f^\mu(\hat{\theta}) - \min_{\eta \in \mathbb{R}^d} f^\mu(\eta) \leqslant \frac{8(1 + \frac{1}{a})G^2 R^2 (32 + \log \frac{1}{\delta})}{\mu n}$$

- Results from Sridharan, Srebro, and Shalev-Shwartz (2008)
  - see also Boucheron and Massart (2011) and references therein

- **Strongly convex functions $\Rightarrow$ fast rate**
  - Warning: μ 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. **Classical stochastic approximation**

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

Outline - II

4. **Non-smooth stochastic approximation**
    - Stochastic (sub)gradient and averaging
    - Non-asymptotic results and lower bounds
    - Strongly convex vs. non-strongly convex

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

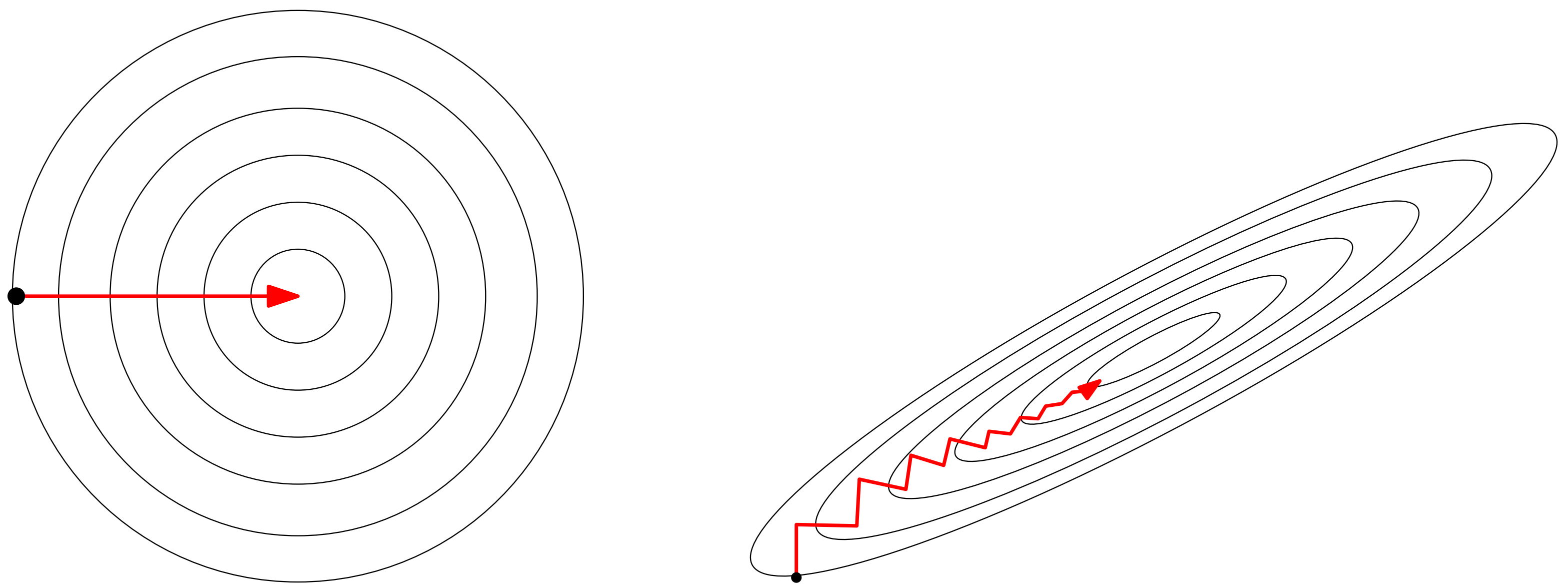
- **Assumption**: g convex on $\mathbb{R}^d$
- **Classical generic algorithms**
  - Gradient descent and accelerated gradient descent
  - Newton method
  - Subgradient method and ellipsoid algorithm
- **Key additional properties of g**
  - Lipschitz continuity, smoothness or strong convexity
- **Key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

(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 μ-strongly convex

- **Algorithm**:

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

- **Bound**:

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

- 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 θ_*

- **Algorithm**:

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

- **Bound**:

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

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

- Gradient descent:

$$\begin{aligned}
\theta_t &= \theta_{t-1} - \frac{1}{L}(H\theta - c) = \theta_{t-1} - \frac{1}{L}(H\theta - H\theta_*) \\
\theta_t - \theta_* &= (I - \frac{1}{L}H)(\theta_{t-1} - \theta_*) = (I - \frac{1}{L}H)^t(\theta_0 - \theta_*)
\end{aligned}$$

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

Gradient descent - Proof for quadratic functions

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

- Gradient descent:

$$\begin{aligned}\theta_t &= \theta_{t-1} - \frac{1}{L}(H\theta - c) = \theta_{t-1} - \frac{1}{L}(H\theta - H\theta_*) \\ \theta_t - \theta_* &= (I - \frac{1}{L}H)(\theta_{t-1} - \theta_*) = (I - \frac{1}{L}H)^t(\theta_0 - \theta_*)\end{aligned}$$

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

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

Properties of smooth convex functions

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

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

- "Distance" to optimum: $g(\theta) - g(\theta_*) \leqslant g'(\theta)^\top(\theta - \theta_*)$

Proof of co-coercivity

- Quadratic upper-bound: $0 \leqslant g(\theta) - g(\eta) - g'(\eta)^\top(\theta - \eta) \leqslant \frac{L}{2}\|\theta - \eta\|^2$
  - Taylor expansion with integral remainder

- Lower bound: $g(\theta) \geqslant g(\eta) + g'(\eta)^\top(\theta - \eta) + \frac{1}{2L}\|g'(\theta) - g'(\eta)\|^2$
  - Define $h(\theta) = g(\theta) - \theta^\top g'(\eta)$, convex with global minimum at η
  - $h(\eta) \leqslant h(\theta - \frac{1}{L}h'(\theta)) \leqslant h(\theta) + h'(\theta)^\top(-\frac{1}{L}h'(\theta))) + \frac{L}{2}\| - \frac{1}{L}h'(\theta))\|^2$, which is thus less than $h(\theta) - \frac{1}{2L}\|h'(\theta)\|^2$
  - Thus $g(\eta) - \eta^\top g'(\eta) \leqslant g(\theta) - \theta^\top g'(\eta) - \frac{1}{2L}\|g'(\theta) - g'(\eta)\|^2$

- Proof of co-coercivity
  - Apply lower bound twice for (η, θ) and (θ, η), and sum to get $0 \geqslant [g'(\eta) - g'(\theta)]^\top(\theta - \eta) + \frac{1}{L}\|g'(\theta) - g'(\eta)\|^2$

- NB: simple proofs with second-order derivatives

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

- Define $h(\theta) = g(\theta) - \theta^\top g'(\eta)$, convex with global minimum at η

- $h(\eta) = \min_\theta h(\theta) \geqslant \min_\zeta h(\theta) + h'(\theta)^\top(\zeta - \theta) + \frac{\mu}{2}\|\zeta - \theta\|^2$, which is attained for $\zeta - \theta = -\frac{1}{\mu}h'(\theta)$
  - This leads to $h(\eta) \geqslant h(\theta) - \frac{1}{2\mu}\|h'(\theta)\|^2$
  - Hence, $g(\eta) - \eta^\top g'(\eta) \geqslant g(\theta) - \theta^\top g'(\eta) - \frac{1}{2\mu}\|g'(\eta) -' (\theta)\|^2$
  - NB: no need for smooothness

- NB: simple proofs with second-order derivatives

- With $\eta = \theta_*$ global minimizer, another "distance" to optimum

$$g(\theta) - g(\theta_*) \leqslant \frac{1}{2\mu}\|g'(\theta)\|^2$$

Convergence proof - gradient descent smooth strongly convex functions

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

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

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

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

Convergence proof - gradient descent smooth convex functions - I

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

$$
\begin{aligned}
\|\theta_t - \theta_*\|^2 &= \|\theta_{t-1} - \theta_* - \gamma g'(\theta_{t-1})\|^2 \\
&= \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|g'(\theta_{t-1})\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\
&\leqslant \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|g'(\theta_{t-1})\|^2 - 2\frac{\gamma}{L}\|g'(\theta_{t-1})\|^2 \text{ using co-coercivity} \\
&= \|\theta_{t-1} - \theta_*\|^2 - \gamma(2/L - \gamma)\|g'(\theta_{t-1})\|^2 \leqslant \|\theta_{t-1} - \theta_*\|^2 \text{ if } \gamma \leqslant 2/L \\
&\leqslant \|\theta_0 - \theta_*\|^2 : \text{ bounded iterates} \\
g(\theta_t) &\leqslant g(\theta_{t-1}) - \frac{1}{2L}\|g'(\theta_{t-1})\|^2 \text{ (see previous slide)} \\
g(\theta_{t-1}) - g(\theta_*) &\leqslant g'(\theta_{t-1})^\top(\theta_{t-1} - \theta_*) \leqslant \|g'(\theta_{t-1})\| \cdot \|\theta_{t-1} - \theta_*\| \text{ (Cauchy-Schwarz)} \\
g(\theta_t) - g(\theta_*) &\leqslant g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L\|\theta_0 - \theta_*\|^2}\big[g(\theta_{t-1}) - g(\theta_*)\big]^2
\end{aligned}
$$

Convergence proof - gradient descent smooth convex functions - II

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

$$\begin{aligned}
g(\theta_t) - g(\theta_*) &\leqslant g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L\|\theta_0 - \theta_*\|^2}\big[g(\theta_{t-1}) - g(\theta_*)\big]^2 \\
\text{of the form } \Delta_k &\leqslant \Delta_{k-1} - \alpha \Delta_{k-1}^2 \text{ with } 0 \leqslant \Delta_k = g(\theta_k) - g(\theta_*) \leqslant \frac{L}{2}\|\theta_k - \theta_*\|^2 \\
\frac{1}{\Delta_{k-1}} &\leqslant \frac{1}{\Delta_k} - \alpha \frac{\Delta_{k-1}}{\Delta_k} \text{ by dividing by } \Delta_k \Delta_{k-1} \\
\frac{1}{\Delta_{k-1}} &\leqslant \frac{1}{\Delta_k} - \alpha \text{ because } (\Delta_k) \text{ is non-increasing} \\
\frac{1}{\Delta_0} &\leqslant \frac{1}{\Delta_t} - \alpha t \text{ by summing from } k = 1 \text{ to } t \\
\Delta_t &\leqslant \frac{\Delta_0}{1 + \alpha t \Delta_0} \text{ by inverting} \\
&\leqslant \frac{2L\|\theta_0 - \theta_*\|^2}{t + 4} \text{ since } \Delta_0 \leqslant \frac{L}{2}\|\theta_k - \theta_*\|^2 \text{ and } \alpha = \frac{1}{2L\|\theta_0 - \theta_*\|^2}
\end{aligned}$$

Limits on convergence rate of first-order methods

- **First-order method**: any iterative algorithm that selects θ_t in $\theta_0 + \mathrm{span}(f'(\theta_0), \dots, f'(\theta_{t-1}))$

- **Problem class**: convex L-smooth functions with a global minimizer θ_*

- **Theorem**: for every integer $k \leqslant (d-1)/2$ and every θ_0, there exist functions in the problem class such that for any first-order method,

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

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

Limits on convergence rate of first-order methods
Proof sketch

- Define quadratic function

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

  - 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 θ_*

- **Algorithm**:

$$\begin{aligned} \theta_t &= \eta_{t-1} - \frac{1}{L} g'(\eta_{t-1}) \\ \eta_t &= \theta_t + \frac{t-1}{t+2}(\theta_t - \theta_{t-1}) \end{aligned}$$

- **Bound**:

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

- 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 θ_*
  - g μ-strongly convex

- **Algorithm**:

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

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

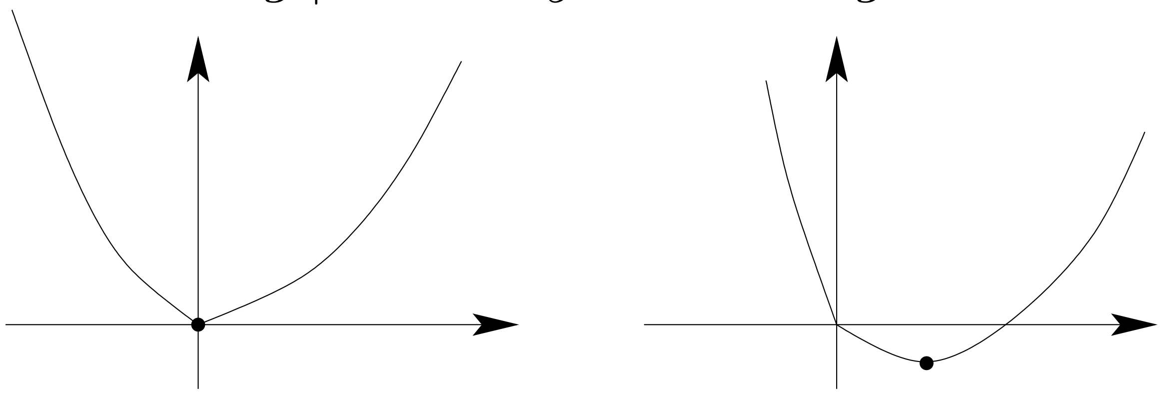
- Gradient descent as a **proximal method** (differentiable functions)
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2}\|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \theta_t - \frac{1}{L}\nabla f(\theta_t)$

- Problems of the form: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) + \mu\Omega(\theta)}$
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \mu\Omega(\theta) + \frac{L}{2}\|\theta - \theta_t\|_2^2$
  - $\Omega(\theta) = \|\theta\|_1 \Rightarrow$ **Thresholded gradient descent**

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

Soft-thresholding for the ℓ_1-norm

- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x\in\mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
  - Derivative at $0+$: $g_+ = \lambda - y$ and $0-$: $g_- = -\lambda - y$

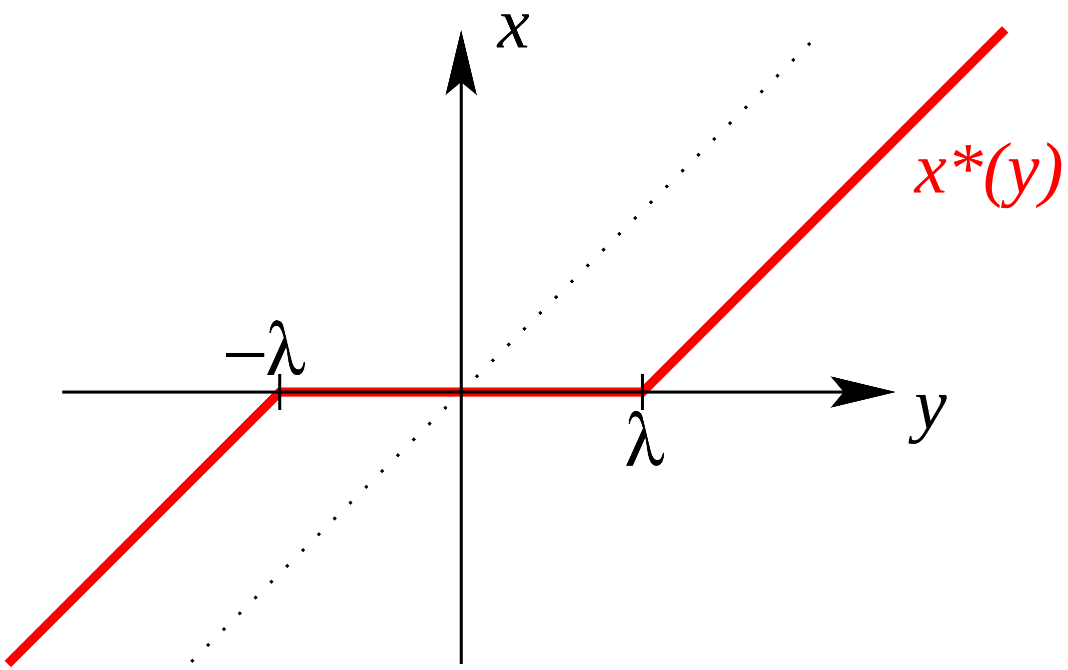


  - $x = 0$ is the solution iff $g_+ \geqslant 0$ and $g_- \leqslant 0$ (i.e., $|y| \leqslant \lambda$)
  - $x \geqslant 0$ is the solution iff $g_+ \leqslant 0$ (i.e., $y \geqslant \lambda$) $\Rightarrow x^* = y - \lambda$
  - $x \leqslant 0$ is the solution iff $g_- \leqslant 0$ (i.e., $y \leqslant -\lambda$)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

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

Soft-thresholding for the ℓ_1-norm

- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
- Solution $\boxed{x^* = \text{sign}(y)(|y| - \lambda)_+}$ = soft thresholding



Newton method

- Given θ_{t-1}, minimize second-order Taylor expansion

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

- **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**: $\theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ 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 $O(e^{-t\sqrt{\mu/L}})$

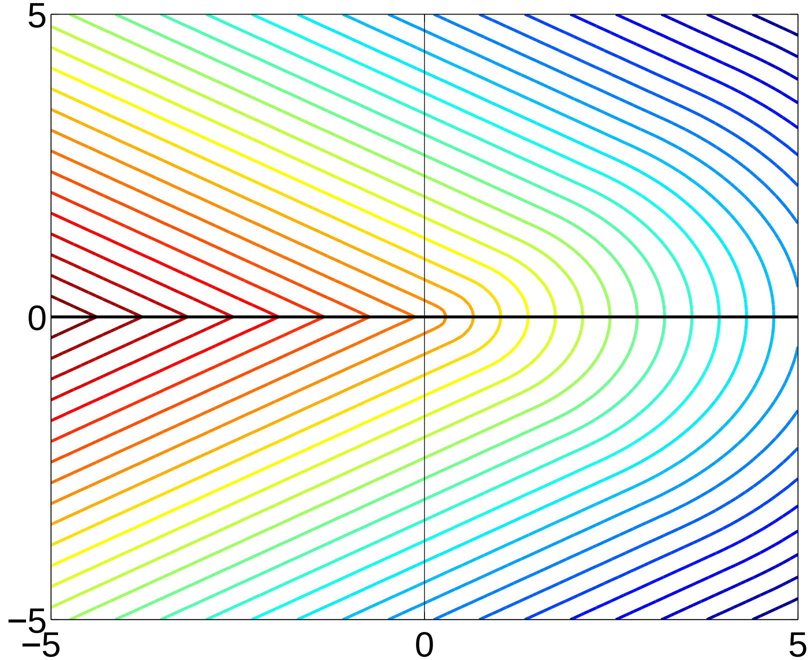
- **Newton method**: $\theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ convergence rate

- **From smooth to non-smooth**
  - Subgradient method and ellipsoid

Counter-example (Bertsekas, 1999)
Steepest descent for nonsmooth objectives

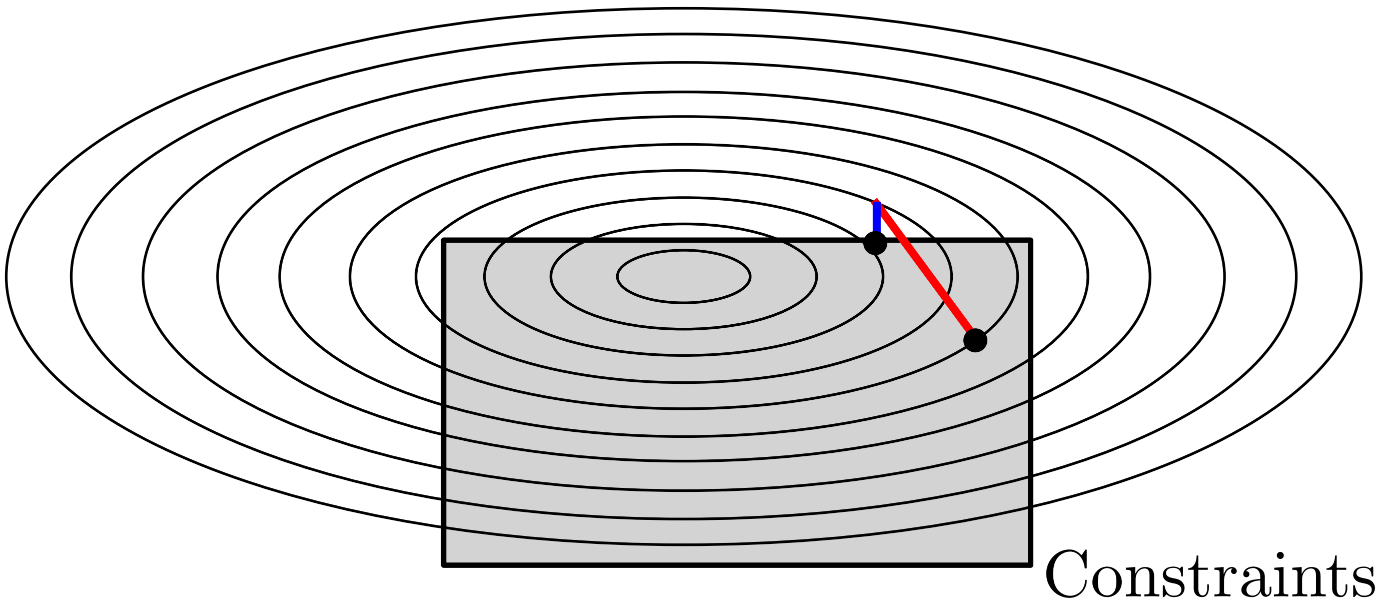
- $g(\theta_1,\theta_2) = \begin{cases} -5(9\theta_1^2+16\theta_2^2)^{1/2} & \text{if } \theta_1 > |\theta_2| \\ -(9\theta_1+16|\theta_2|)^{1/2} & \text{if } \theta_1 \leqslant |\theta_2| \end{cases}$

- Steepest descent starting from any θ 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 $\{\|\theta\|_2 \leqslant D\}$
- **Algorithm**: $\theta_t = \Pi_D\left(\theta_{t-1} - \frac{2D}{B\sqrt{t}} g'(\theta_{t-1})\right)$
  - Π_D : orthogonal projection onto $\{\|\theta\|_2 \leqslant D\}$



Subgradient method/"descent" (Shor et al., 1985)

- **Assumptions**
  - g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
- **Algorithm**: $\theta_t = \Pi_D\left(\theta_{t-1} - \frac{2D}{B\sqrt{t}} g'(\theta_{t-1})\right)$
  - Π_D : orthogonal projection onto $\{\|\theta\|_2 \leqslant D\}$
- **Bound**:

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

- 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: $\|g'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$

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

- leading to

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

Subgradient method/"descent" - proof - II

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

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

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

- Optimized step-size $\gamma_t = \frac{2D}{B\sqrt{t}}$ depends on "horizon"
  - Leads to bound of $2DB\sqrt{t}$

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

Subgradient method/"descent" - proof - III

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

- Decreasing step-size

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

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

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 $\|\Phi(x)\|_2 \leqslant R$ 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 $\Theta = \{\|\theta\|_2 \leqslant D\}$

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

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

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

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

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

Subgradient descent - strong convexity

- **Assumptions**
  - g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
  - g μ-strongly convex

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

- **Bound**:

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

- 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: $\|g'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$ and μ-strong convexity of f

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

- leading to

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

Subgradient method - strong convexity - proof - II

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

$$\sum_{u=1}^{t} u\big[g(\theta_{u-1}) - g(\theta_*)\big] \leqslant \sum_{t=1}^{u} \frac{B^2 u}{\mu(u+1)} + \frac{1}{4}\sum_{u=1}^{t}\big[u(u-1)\|\theta_{u-1} - \theta_*\|_2^2 - u(u+1)\|\theta_u - \theta_*\|_2^2\big]$$

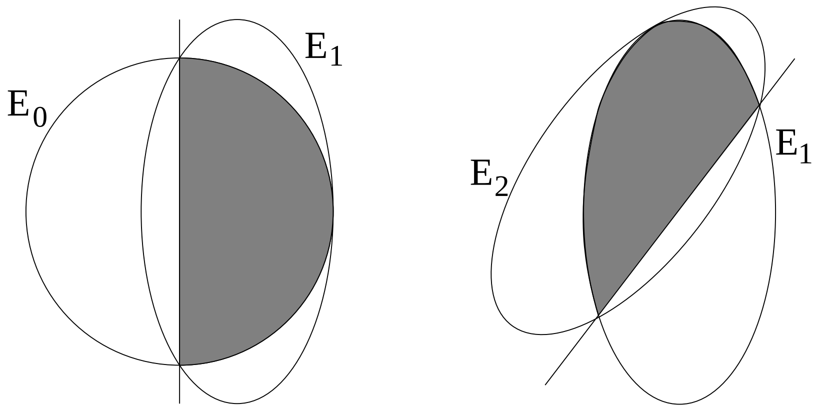
$$\leqslant \frac{B^2 t}{\mu} + \frac{1}{4}\big[0 - t(t+1)\|\theta_t - \theta_*\|_2^2\big] \leqslant \frac{B^2 t}{\mu}$$

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

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

Ellipsoid method

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



- Represent $E_t = \{\theta \in \mathbb{R}^d, (\theta - \theta_t)^\top P_t^{-1}(\theta - \theta_t) \leqslant 1\}$

- 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 $h_t = \frac{1}{\sqrt{g'(\theta_t)^\top P_t g'(x_t)}} g'(\theta_t)$

- Fact 2: $\mathrm{vol}(\mathcal{E}_t) \approx \mathrm{vol}(\mathcal{E}_{t-1}) e^{-1/2d} \Rightarrow$ 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**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ 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**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ 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
  - μ 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. **Classical stochastic approximation**
    - Asymptotic analysis
    - Robbins-Monro algorithm
    - Polyak-Rupert averaging

Outline - II

4. **Non-smooth stochastic approximation**

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

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

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

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

"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, 1951)**

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

- **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

"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 $\theta_n = \frac{2}{n(n+1)}\sum_{k=1}^n k x_k$

"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

Decomposition of the noise term

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

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

Decomposition of the noise term

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

- Require γ_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 for recursive mean estimation: need $\alpha \in (0,1)$

Robbins-Monro algorithm

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

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

- **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 $\theta_n \to \theta_*$ or $d(\theta_n, \Theta_*) \to 0$ or $f(\theta_n) \to f(\theta_*)$
  – Random quantity $\delta_n \in \mathbb{R}$ tending to zero

- **Convergence almost-surely**: $\mathbb{P}(\delta_n \to 0) = 1$

- **Convergence in probability**: $\forall \varepsilon > 0, \mathbb{P}(|\delta_n| \geqslant \varepsilon) \to 0$

- **Convergence in mean** $r \geqslant 1$: $\mathbb{E}|\delta_n|^r \to 0$

Different types of convergences

- **Goal**: show that $\theta_n \to \theta_*$ or $d(\theta_n, \Theta_*) \to 0$ or $f(\theta_n) \to f(\theta_*)$
  - Random quantity $\delta_n \in \mathbb{R}$ tending to zero

- **Convergence almost-surely**: $\mathbb{P}(\delta_n \to 0) = 1$

- **Convergence in probability**: $\forall \varepsilon > 0, \mathbb{P}(|\delta_n| \geqslant \varepsilon) \to 0$

- **Convergence in mean** $r \geqslant 1$: $\mathbb{E}|\delta_n|^r \to 0$

- **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)

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

- The Robbins-Monro algorithm cannot converge all the time...

- **Lyapunov function** $V : \mathbb{R}^d \to \mathbb{R}$ with following properties
  - Non-negative values: $V \geqslant 0$
  - Continuously-differentiable with L-Lipschitz-continuous gradients
  - Control of h: $\forall \theta$, $\|h(\theta)\|^2 \leqslant C(1 + V(\theta))$
  - Gradient condition: $\forall \theta$, $\boxed{h(\theta)^\top V'(\theta) \geqslant \alpha \|V'(\theta)\|^2}$

Robbins-Monro algorithm
Need for Lyapunov functions (even with no noise)

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

- The Robbins-Monro algorithm cannot converge all the time...

- **Lyapunov function** $V : \mathbb{R}^d \to \mathbb{R}$ with following properties
  - Non-negative values: $V \geqslant 0$
  - Continuously-differentiable with L-Lipschitz-continuous gradients
  - Control of h: $\forall \theta$, $\|h(\theta)\|^2 \leqslant C(1 + V(\theta))$
  - Gradient condition: $\forall \theta$, $\boxed{h(\theta)^\top V'(\theta) \geqslant \alpha \|V'(\theta)\|^2}$

- 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 $\|V'(\theta)\|^2 \geqslant 2\mu V(\theta)$

Robbins-Monro algorithm
Martingale noise

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

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

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

Robbins-Monro algorithm
Convergence of the Lyapunov function

- Using regularity (and other properties) of V:

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

Robbins-Monro algorithm
Convergence of the expected Lyapunov function
with “curvature”

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

$$\begin{aligned}\mathbb{E}\big[V(\theta_n)|\mathcal{F}_{n-1}\big] &\leqslant V(\theta_{n-1})\big[1-\alpha\mu\gamma_n\big] + M\gamma_n^2 \\ \mathbb{E}V(\theta_n) &\leqslant \mathbb{E}V(\theta_{n-1})\big[1-\alpha\mu\gamma_n\big] + M\gamma_n^2\end{aligned}$$

- Need to study non-negative sequence $\delta_n \leqslant \delta_{n-1}\big[1-\mu\gamma_n\big] + M\gamma_n^2$ 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 $\gamma_n \to 0$
  - 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 $\delta_n \leqslant \delta_{n-1}\big[1 - \mu\gamma_n\big] + M\gamma_n^2$ with $\delta_n = \mathbb{E}V(\theta_n)$ (NB: forgetting constraint on γ_n - see next class)

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

- If $\alpha > 1$: no forgetting of initial conditions

- If $\alpha \in (0,1)$: $\delta_0 \exp(-\text{ cst } \mu C \times n^{\alpha-1}) + \gamma_n M$

- If $\alpha = 1$ and $\gamma_n = C/n$: $\delta_0 n^{-\mu C} + \gamma_n M$

Robbins-Monro algorithm
Almost-sure convergence

- Using regularity of V:

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

Robbins and Siegmund (1985)

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

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

- *Proof*

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

Robbins and Siegmund (1985) - Proof

- Inequality: $\mathbb{E}\big[V_n|\mathcal{F}_{n-1}\big] \leqslant V_{n-1}(1+\beta_{n-1}) + \chi_{n-1} - \eta_{n-1}$
- 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:

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

- Define the super-martingale $Y_n = V'_n - \sum_{k=1}^{n-1}(\chi'_k - \eta'_k)$ so that $\mathbb{E}\big[Y_n|\mathcal{F}_{n-1}\big] \leqslant Y_{n-1}$
- Deterministic proof
- 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 θ_* symmetric)

$$
\begin{aligned}
\theta_n &= \theta_{n-1} - \gamma_n h(\theta_{n-1}) - \gamma_n \varepsilon_n \\
&\approx \theta_{n-1} - \gamma_n \big[h'(\theta_*)(\theta_{n-1} - \theta_*)\big] - \gamma_n \varepsilon_n + \gamma_n O(\|\theta_n - \theta_*\|^2) \\
&\approx \theta_{n-1} - \gamma_n A(\theta_{n-1} - \theta_*) - \gamma_n \varepsilon_n \\
\theta_n - \theta_* &\approx (I - \gamma_n A) \cdots (I - \gamma_1 A)(\theta_0 - \theta_*) - \sum_{k=1}^{n} (I - \gamma_n A) \cdots (I - \gamma_{k+1} A) \gamma_k \varepsilon_k \\
\theta_n - \theta_* &\approx \exp\big[-(\gamma_n + \cdots + \gamma_1)A\big](\theta_0 - \theta_*) - \sum_{k=1}^{n} \exp\big[-(\gamma_n + \cdots + \gamma_{k+1})A\big] \gamma_k \varepsilon_k \\
&\approx \exp\big[-CA\log n\big](\theta_0 - \theta_*) - \sum_{k=1}^{n} \exp\big[-C(\log n - \log k)A\big] \frac{C}{k} \varepsilon_k
\end{aligned}
$$

- 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

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

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

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 $2C\lambda_{\min}(A) \geqslant 1$ 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 $\boxed{\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_n}$
  - 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 θ_*. See details by Polyak and Juditsky (1992)

Cesaro means

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

Cesaro means

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

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

  - Will depend on rate α_n
  - If $\sum_n \alpha_n < \infty$, the rate becomes $1/n$ independently of α_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 $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$

$$
\begin{aligned}
h(\theta_{n-1}) &= \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n \\
A(\theta_{n-1} - \theta_*) + O(\|\theta_{n-1} - \theta_*\|^2) &= \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n \text{ with } A = h'(\theta_*) \\
A(\theta_{n-1} - \theta_*) &= \frac{1}{\gamma_n}\big[\theta_{n-1} - \theta_n\big] - \varepsilon_n + O(n^{-\alpha}) \\
\frac{1}{n}\sum_{k=1}^n A(\theta_{k-1} - \theta_*) &= \frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}\big[\theta_{k-1} - \theta_k\big] - \frac{1}{n}\sum_{k=1}^n \varepsilon_k + O(n^{-\alpha}) \\
\frac{1}{n}\sum_{k=1}^n A(\theta_{k-1} - \theta_*) &= \frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}\big[\theta_{k-1} - \theta_k\big] + \text{Normal}(0, \Sigma/n) + O(n^{-\alpha})
\end{aligned}
$$

Polyak-Ruppert averaging - Proof sketch - II

- **Goal**: Bounding $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\big[\theta_{k-1}-\theta_k\big]$ 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

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

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 $\|\theta_n - \theta_*\|^2 = O(n^{-\alpha})$

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

- **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. **Classical stochastic approximation**

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

Outline - II

4. **Non-smooth stochastic approximation**
   - Stochastic (sub)gradient and averaging
   - Non-asymptotic results and lower bounds
   - Strongly convex vs. non-strongly convex

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**: $\boxed{f_n(\theta) = \ell(y_n, \theta^\top \Phi(x_n))}$
  - $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E}\,\ell(y_n, \theta^\top \Phi(x_n))$ = **generalization error**
  - Expected gradient: $f'(\theta) = \mathbb{E} f'_n(\theta) = \mathbb{E}\big\{\ell'(y_n, \theta^\top \Phi(x_n))\,\Phi(x_n)\big\}$
  - 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

$$\theta_n = \theta_{n-1} - \gamma_n h_n(\theta_{n-1}) \text{ with } \mathbb{E}\big[h_n(\theta_{n-1})|\theta_{n-1}\big] = h(\theta_{n-1})$$

  - Beyond convex problems, i.i.d assumption, finite dimension, etc.
  - Typically asymptotic results
  - 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 θ
  - 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 θ
  - 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 Θ
  - Generalization bound using uniform concentration results

Relationship to online learning

- **Stochastic approximation**
  - Minimize $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$ **generalization error** of θ
  - 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 Θ
  - 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 μ-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 μ-strongly convex

- **Key algorithm:** Stochastic gradient descent (a.k.a. Robbins-Monro)

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

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1} \theta_k$
  - Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = Cn^{-\alpha}}$

Convex stochastic approximation

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

- **Key algorithm:** Stochastic gradient descent (a.k.a. Robbins-Monro)

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

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$
  - Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = Cn^{-\alpha}}$

- **Desirable practical behavior**
  - Applicable (at least) to classical supervised learning problems
  - Robustness to (potentially unknown) constants (L,B,μ)
  - Adaptivity to difficulty of the problem (e.g., strong convexity)

Stochastic subgradient "descent"/method

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

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

Stochastic subgradient "descent"/method

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

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

- **Bound**:

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

- "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
- $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, unbiased gradients/functions $\mathbb{E}(f_n|\mathcal{F}_{n-1}) = f$

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

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

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

Stochastic subgradient method - proof - II

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

$$\sum_{u=1}^{n}\big[\mathbb{E} f(\theta_{u-1}) - f(\theta_*)\big] \leqslant \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{n}\frac{1}{2\gamma_u}\big[\mathbb{E}\|\theta_{u-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_u - \theta_*\|_2^2\big]$$

$$\leqslant \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} \leqslant 2DB\sqrt{n} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}}$$

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

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 $\| f_n'(\theta_{n-1}) + \varepsilon_n \| \leqslant B$ almost surely

- **Performance criterion**: (normalized) regret

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

  - 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 - θ an arbitrary point such that $\|\theta\| \leqslant D$

- $\|f'_n(\theta_{n-1}) + \varepsilon_n\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, unbiased gradients $\mathbb{E}(\varepsilon_n|\mathcal{F}_{n-1}) = 0$

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

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

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

Stochastic subgradient method - online learning - II

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

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

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

- 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 μ-strongly convex on $\{\|\theta\|_2 \leqslant D\}$
  - θ_* global optimum of f over $\{\|\theta\|_2 \leqslant D\}$

- **Algorithm**: $\theta_n = \Pi_D\Big(\theta_{n-1} - \frac{2}{\mu(n+1)} f_n'(\theta_{n-1})\Big)$

- **Bound**:

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

- "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: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{t-1}))$ with $\gamma_n = \frac{2}{\mu(n+1)}$

- Assumption: $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$ and μ-strong convexity of f

$$\begin{aligned}
\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\
&\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f_n'(\theta_{t-1}) \text{ because } \|f_n'(\theta_{t-1})\|_2 \leqslant B \\
\mathbb{E}(\cdot|\mathcal{F}_{n-1}) &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2}\|\theta_{n-1} - \theta_*\|_2^2\big]
\end{aligned}$$

(property of subgradients and strong convexity)

- leading to

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

Stochastic subgradient - strong convexity - proof - II

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

$$\sum_{u=1}^n u\big[\mathbb{E}f(\theta_{u-1}) - f(\theta_*)\big] \leqslant \sum_{u=1}^n \frac{B^2 u}{\mu(u+1)} + \frac{1}{4}\sum_{u=1}^n \big[u(u-1)\mathbb{E}\|\theta_{u-1}-\theta_*\|_2^2 - u(u+1)\mathbb{E}\|\theta_u - \theta_*$$

$$\leqslant \quad \frac{B^2 n}{\mu} + \frac{1}{4}\big[0 - n(n+1)\mathbb{E}\|\theta_n - \theta_*\|_2^2\big] \leqslant \frac{B^2 n}{\mu}$$

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

- 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 $\mathbb{E} f_n = f$
  - θ_* global optimum of $g = f + \frac{\mu}{2}\| \cdot \|_2^2$
  - No compactness assumption - no projections

- **Algorithm**:

$$\theta_n = \theta_{n-1} - \frac{2}{\mu(n+1)} g_n'(\theta_{n-1}) = \theta_{n-1} - \frac{2}{\mu(n+1)} \big[f_n'(\theta_{n-1}) + \mu \theta_{n-1} \big]$$

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

- Minimax convergence rate

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

- Iteration: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{t-1}))$ with $\gamma_n = \frac{1}{\mu n}$

- Assumption: $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$ and μ-strong convexity of f

$$\begin{aligned}
\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\
&\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f_n'(\theta_{t-1}) \text{ because } \|f_n'(\theta_{t-1})\|_2 \leqslant B \\
\mathbb{E}(\cdot|\mathcal{F}_{n-1}) &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2}\|\theta_{n-1} - \theta_*\|_2^2\big] \\
&\quad \text{(property of subgradients and strong convexity)}
\end{aligned}$$

- leading to

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

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

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

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

- Using convexity: $\mathbb{E} f\bigg(\dfrac{1}{n}\displaystyle\sum_{u=1}^n \theta_{u-1}\bigg) - f(\theta_*) \leqslant \dfrac{B^2}{2\mu}\dfrac{\log n}{n}$

- Why could this be useful?

Stochastic subgradient descent - strong convexity
Online learning

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

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

- Not optimal. See Hazan and Kale (2014).

Beyond convergence in expectation

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

- **High-probability bounds**
  - Markov inequality: $\mathbb{P}\Big(f\Big(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\Big) - f(\theta_*) \geqslant \varepsilon\Big) \leqslant \frac{2DB}{\sqrt{n}\varepsilon}$

Beyond convergence in expectation

- **Typical result**: $\mathbb{E}f\bigg(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\bigg) - f(\theta_*) \leqslant \frac{2DB}{\sqrt{n}}$
  - Obtained with simple conditioning arguments

- **High-probability bounds**
  - Markov inequality: $\mathbb{P}\Big(f\Big(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\Big) - f(\theta_*) \geqslant \varepsilon\Big) \leqslant \frac{2DB}{\sqrt{n}\varepsilon}$
  - Concentration inequality (Nemirovski et al., 2009; Nesterov and Vial, 2008)

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

- 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

- $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, unbiased gradients/functions $\mathbb{E}(f_n|\mathcal{F}_{n-1}) = f$

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

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

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

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

Stochastic subgradient method - high probability - II

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

$$\begin{aligned} \|\theta_n-\theta_*\|_2^2 &\leqslant \|\theta_{n-1}-\theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1})-f(\theta_*)\big] + Z_n \\ f(\theta_{n-1})-f(\theta_*) &\leqslant \frac{1}{2\gamma_n}\big[\|\theta_{n-1}-\theta_*\|_2^2 - \|\theta_n-\theta_*\|_2^2\big] + \frac{B^2\gamma_n}{2} + \frac{Z_n}{2\gamma_n} \end{aligned}$$

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

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

Stochastic subgradient method - high probability - III

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

- Azuma-Hoeffding inequality for bounded martingale increments:

$$\mathbb{P}\Big(\sum_{u=1}^{n} \frac{Z_u}{2\gamma_u} \geqslant t\sqrt{n} \cdot 4DB\Big) \leqslant \exp\Big(-\frac{t^2}{2}\Big)$$

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

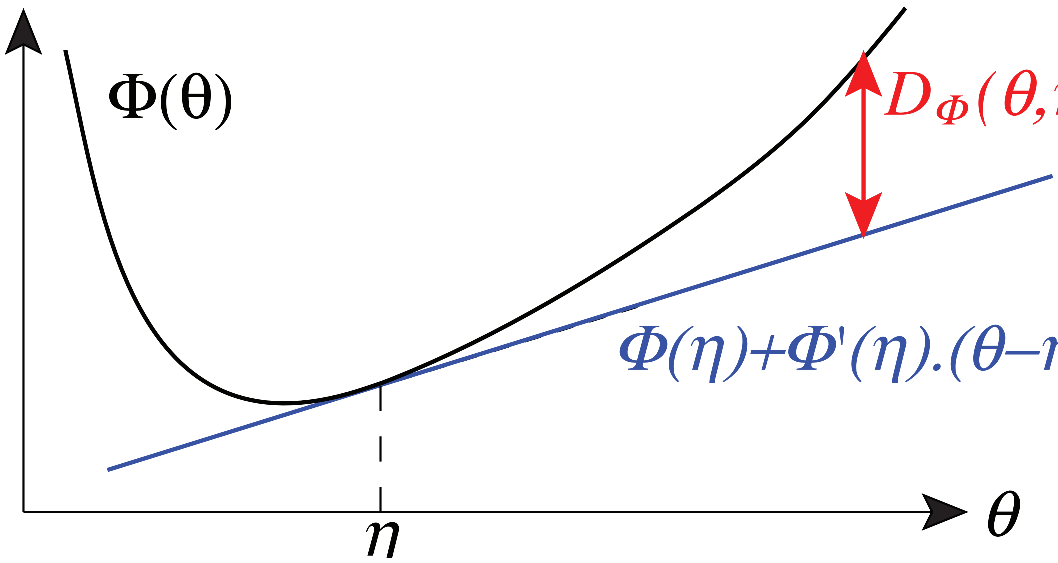
- What about other norms?
  - Example: natural bound on $\max_{\theta\in\Theta}\|f'(\theta)\|_\infty$ leads to $\sqrt{d}$ factor
  - Avoidable with mirror descent, which leads to factor $\sqrt{\log d}$
  - Nemirovski et al. (2009); Lan et al. (2012)

- 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 $\|s\|_* = \sup_{\|\theta\|\leqslant 1} \theta^\top s$
- B-Lipschitz-continuous function w.r.t. $\|\cdot\|$: $\|f'(\theta)\|_* \leqslant B$
- Given a strictly-convex function Φ, define the Bregman divergence

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



Mirror map

- Strongly-convex function $\Phi : \mathcal{C}_\Phi \to \mathbb{R}$ such that

(a) the gradient Φ' takes all possible values in $\mathbb{R}^d$, leading to a bijection from $\mathcal{C}_\Phi$ to $\mathbb{R}^d$

(b) the gradient Φ' 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 $\mathcal{C}_\Phi$:

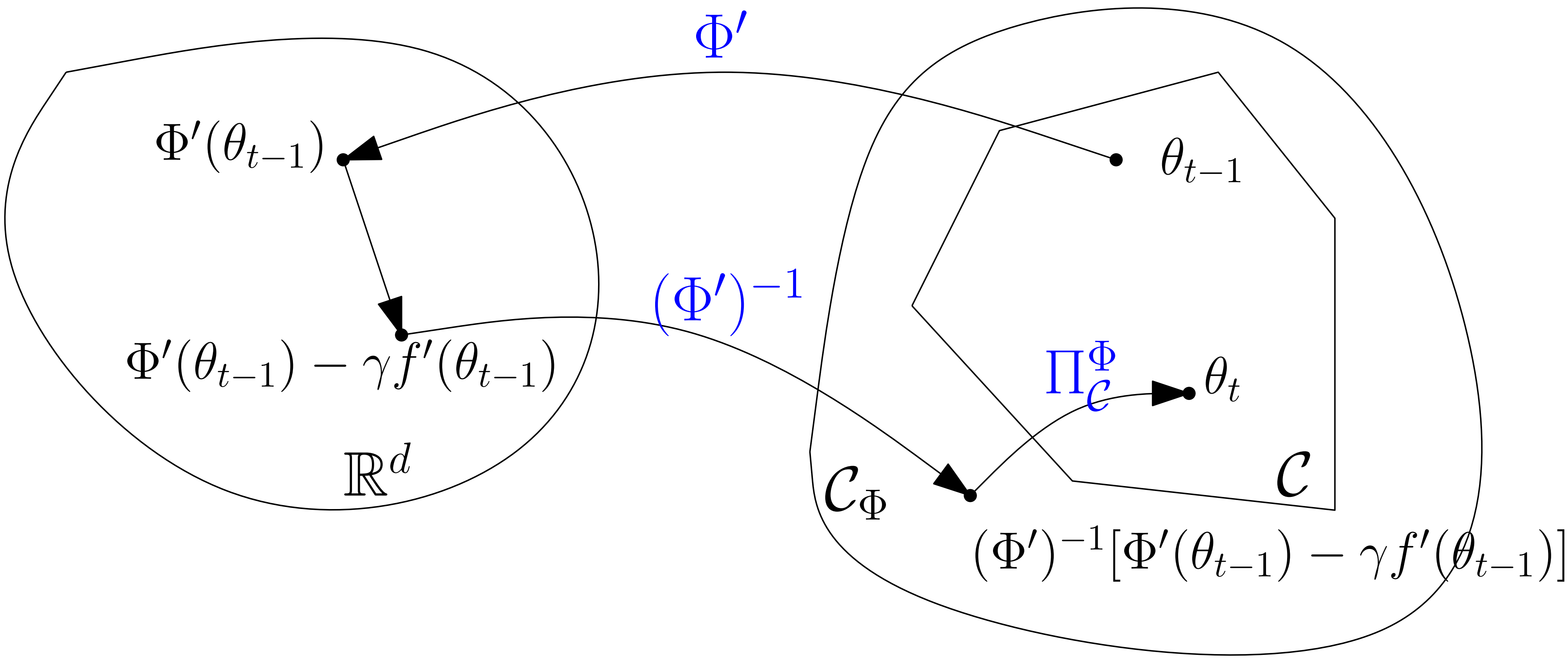
$$\begin{aligned}
\Pi_{\mathcal{C}}^{\Phi}(\theta) &= \arg \min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} D_\Phi(\eta, \theta) \\
&= \arg \min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi(\theta) - \Phi'(\theta)^\top (\eta - \theta) \\
&= \arg \min_{\eta \in \mathcal{C}_\Phi \cap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

- Example of squared Euclidean norm and entropy

Mirror descent

- **Iteration**:

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



Mirror descent

- **Iteration**:

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

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

$$f\bigg(\frac{1}{t}\sum_{u=1}^{t}\theta_u\bigg) \leqslant DB\sqrt{\frac{2}{\alpha t}}$$

  – 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

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

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

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

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

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 $\mathcal{C}_\Phi = \big\{\theta \in \mathbb{R}_+^d,\ \sum_{i=1}^d \theta_i = 1\big\}$
  - Bregman divergence = Kullback-Leibler divergence
  - Iteration (multiplicative update): $\theta_t \propto \theta_{t-1} \exp(-\gamma f'(\theta_{t-1}))$
  - Constant: $D^2 = \log d$, $\alpha = 1$
- **ℓ_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 ℓ_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 θ_n

- **Class of functions**
  - convex B-Lipschitz-continuous (w.r.t. ℓ_2-norm) on a compact convex set $\mathcal{C}$ containing an ℓ_∞-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: $\displaystyle \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$

- **Minimax performance**: $\displaystyle \inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$

Minimax rates (Agarwal et al., 2012)

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

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

  - Consequences for ℓ_2-ball of radius D: $BD/\sqrt{n}$
  - Upper-bound through stochastic subgradient

- μ**-strongly-convex functions**:

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

Minimax rates - sketch of proof

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

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

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 α^j, $j = 1, \dots, M$ of the hypercube $\{-1, 1\}^d$, which are sufficiently far in Hamming metric Δ_H (denote $\mathcal{V}$ this set with $|\mathcal{V}| = M$)
$$\forall j \neq k, \ \Delta_H(\alpha^i, \alpha^j) \geqslant \frac{d}{4},$$
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 α^j, $j = 1, \dots, M$

Minimax rates - sketch of proof

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

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

   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 α^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 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: $\sum_{\alpha \in \mathcal{V}} |\mathcal{B}_H(\alpha, d/4)| \geqslant 2^d$
3. Link with deviation of Z distributed as $\mathrm{Binomial}(d, 1/2)$

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

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

Designing a class of functions

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

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

- **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 $\hat{g}_\alpha(x) = c\big[b_i f_i^+ + (1-b_i)f_i^-\big]$ and its value / gradient

Optimizing is function identification

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

- **"Metric" between functions**:

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

  – $\rho(f,g) \geqslant 0$ 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_α such that $g_\alpha(\tilde{\theta}) - \inf_{\theta \in \mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$

Optimizing is function identification

- **Goal**: if g_α is optimized up to error ε, then this identifies $\alpha \in \mathcal{V}$
- **"Metric" between functions**:

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

  – $\rho(f,g) \geqslant 0$ 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_α such that $g_\alpha(\tilde\theta) - \inf_{\theta\in\mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$
  – (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 α^* 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

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

  - "d-term" comes from $\log m = \log|\mathcal{V}| \propto d$
  - "n-term" comes from n independent draws
  - "δ-term" comes from biasedness of coins proportional to δ

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 ℓ_2-norm. With $c = B/2$, then g_α is B-Lipschitz.

- Lower bound on the discrepancy function
  - each g_α is minimized at $\theta_\alpha = -\alpha/2$
  - Fact: $\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\Delta_H(\alpha,\beta) \geqslant \frac{c\delta}{2} = \psi(\delta)$

- Set error/precision $\varepsilon = \frac{c\delta}{18}$ so that $\varepsilon < \psi(\delta)/9$

- Consequence: $\frac{1}{3} \geqslant 1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}$, that is, $\boxed{n \geqslant \text{ cst } \times \frac{L^2 d^2}{\varepsilon^2}}$

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
  - μ 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. **Classical stochastic approximation**
   - Asymptotic analysis
   - Robbins-Monro algorithm
   - Polyak-Rupert averaging

Outline - II

4. **Non-smooth stochastic approximation**

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

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: $\boxed{\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})}$
  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$
- **Smoothness of f_n**: For each $n \geqslant 1$, 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: $O\Big(\frac{\sigma^2\gamma_n}{\mu}\Big) + O(e^{-\mu n \gamma_n})\|\theta_0 - \theta_*\|^2$
  - averaging: $\frac{\operatorname{tr} H(\theta_*)^{-1}}{n} + \mu^{-1}O(n^{-2\alpha} + n^{-2+\alpha}) + O\Big(\frac{\|\theta_0 - \theta_*\|^2}{\mu^2 n^2}\Big)$

Classical proof sketch (no averaging) - I

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

Classical proof sketch (no averaging) - II

- **Main bound**

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

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

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

Decomposition of the noise term

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

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

Decomposition of the noise term

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

- Require γ_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 for recursive mean estimation: need $\alpha \in (0,1)$

Proof sketch (averaging)

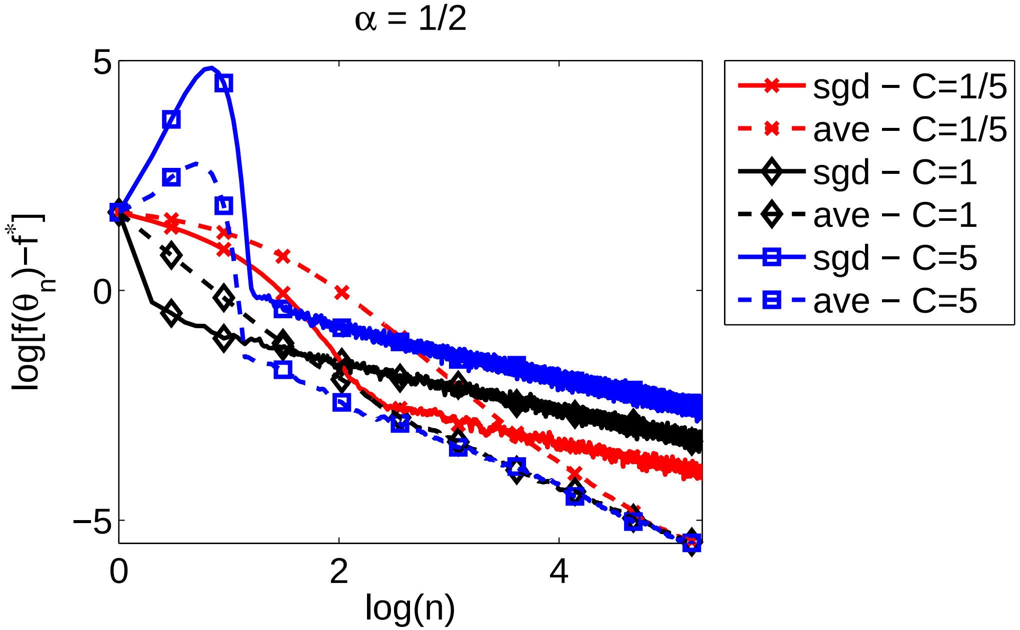
- From Polyak and Juditsky (1992):

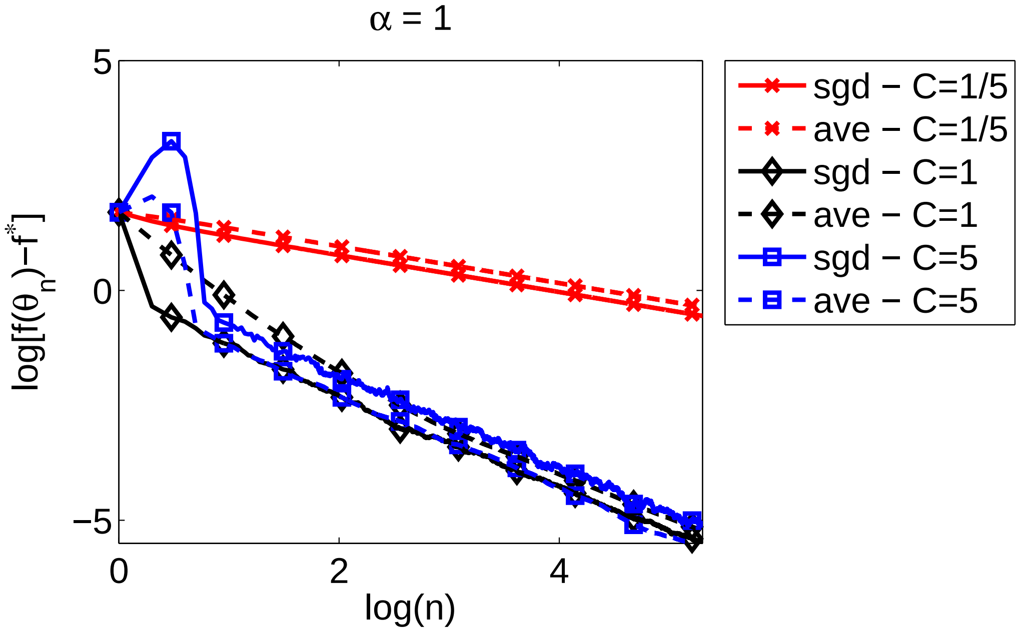
$$
\begin{aligned}
& \theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1}) \\
\Leftrightarrow \quad & f_n'(\theta_{n-1}) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n) \\
\Leftrightarrow \quad & f_n'(\theta_*) + f_n''(\theta_*)(\theta_{n-1} - \theta_*) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n) + O(\|\theta_{n-1} - \theta_*\|^2) \\
\Leftrightarrow \quad & f_n'(\theta_*) + f''(\theta_*)(\theta_{n-1} - \theta_*) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n) + O(\|\theta_{n-1} - \theta_*\|^2) \\
& \qquad + O(\|\theta_{n-1} - \theta_*\|)\varepsilon_n \\
\Leftrightarrow \quad & \theta_{n-1} - \theta_* = -f''(\theta_*)^{-1} f_n'(\theta_*) + \frac{1}{\gamma_n} f''(\theta_*)^{-1}(\theta_{n-1} - \theta_n) \\
& \qquad + O(\|\theta_{n-1} - \theta_*\|^2) + O(\|\theta_{n-1} - \theta_*\|)\varepsilon_n
\end{aligned}
$$

- 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)

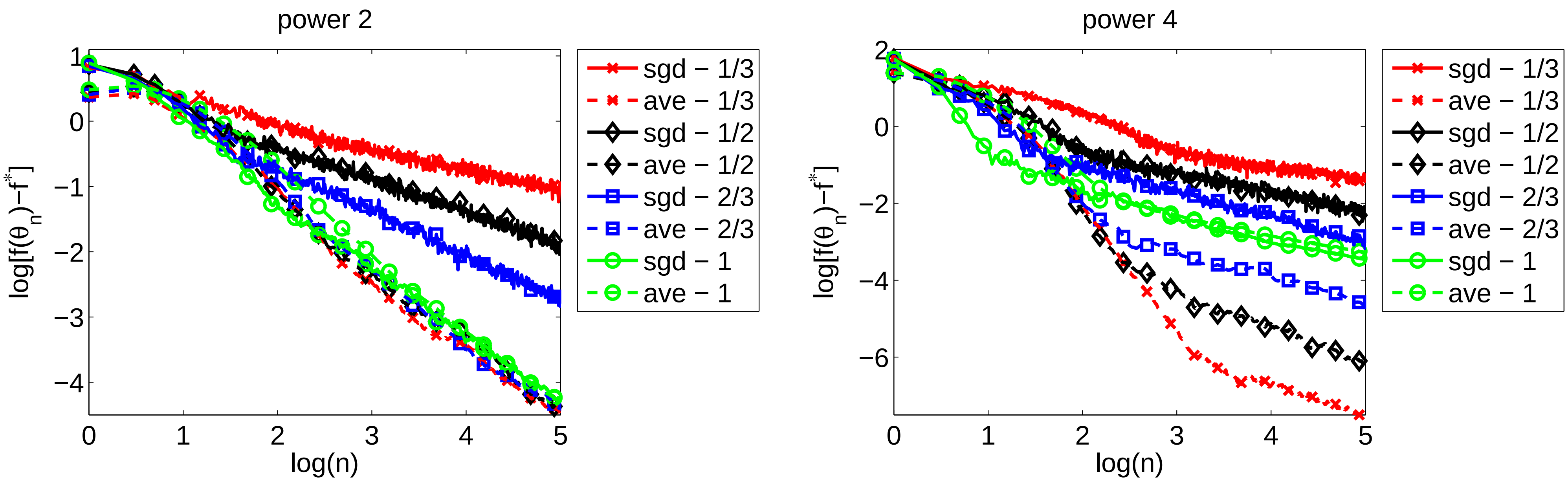
- Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old: $O(n^{-1})$ rate achieved without averaging for $\alpha = 1$
  - New: $O(n^{-1})$ rate achieved with averaging for $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with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

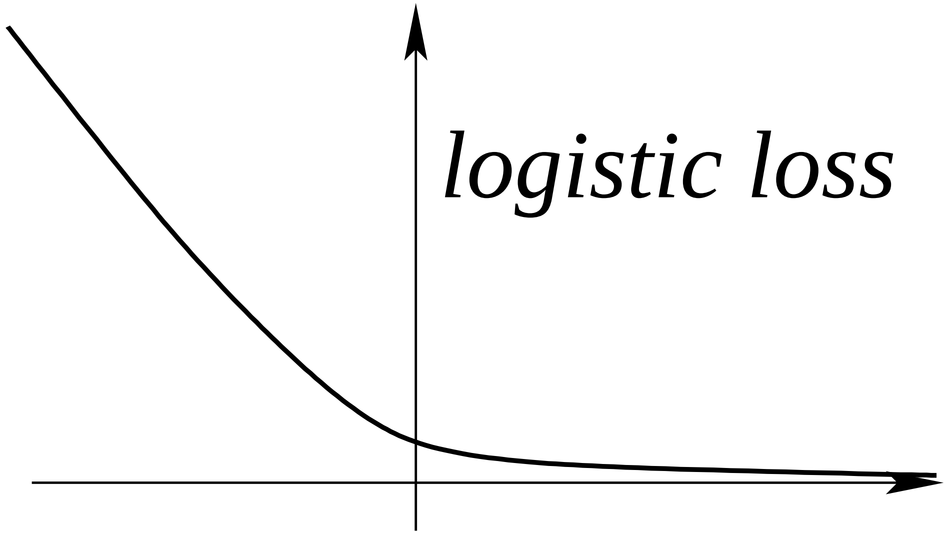
Adaptive algorithm for logistic regression

- **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 $|\theta^\top \Phi(x_n)| \leqslant M$ (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 $|\theta^\top \Phi(x_n)| \leqslant M$ (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** $1/\big(2R^2\sqrt{n}\big)$
  - with $R =$ radius of data (Bach, 2013):

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

  - Proof based on self-concordance (Nesterov and Nemirovski, 1994)

Self-concordance - I

- Usual definition for convex $\varphi : \mathbb{R} \rightarrow \mathbb{R}$: $|\varphi'''(t)| \leqslant 2\varphi''(t)^{3/2}$
  - 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: $|\varphi'''(t)| \leqslant \varphi''(t)$
  - 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 $\varphi : \mathbb{R} \to \mathbb{R}$: $|\varphi'''(t)| \leqslant 2\varphi''(t)^{3/2}$
  - 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: $|\varphi'''(t)| \leqslant \varphi''(t)$
  - Applicable to logistic regression (with extensions)
  - If features bounded by R, $h : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satisfies: $\forall t \in \mathbb{R}$, $|h'''(t)| \leqslant R\|\theta_1 - \theta_2\| h''(t)$

Self-concordance - I

- Usual definition for convex $\varphi : \mathbb{R} \rightarrow \mathbb{R}$: $|\varphi'''(t)| \leqslant 2\varphi''(t)^{3/2}$
  - 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: $|\varphi'''(t)| \leqslant \varphi''(t)$
  - Applicable to logistic regression (with extensions)
  - If features bounded by R, $h : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satisfies: $\forall t \in \mathbb{R}$, $|h'''(t)| \leqslant R\|\theta_1 - \theta_2\| h''(t)$

- **Important properties**
  - Allows global Taylor expansions
  - Relates expansions of derivatives of different orders

Global Taylor expansions

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

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

Global Taylor expansions

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

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

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

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

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

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

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

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

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

Relating Taylor expansions of different orders

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

Relating Taylor expansions of different orders

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

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

$$
\begin{aligned}
\varphi'(t) &= \big\langle z, f''\big(\theta_2 + t(\theta_1 - \theta_2)\big) - f''(\theta_2), \theta_1 - \theta_2\big\rangle \\
\varphi''(t) &= f'''\big(\theta_2 + t(\theta_1 - \theta_2)\big)[z, \theta_1 - \theta_2, \theta_1 - \theta_2] \\
&\leqslant R\|z\|_2 f''\big(\theta_2 + t(\theta_1 - \theta_2)\big)[\theta_1 - \theta_2, \theta_1 - \theta_2], \text{ using App. A of Bach (2010)} \\
&= R\big\langle \theta_2 - \theta_1, f''\big(\theta_2 + t(\theta_1 - \theta_2)\big)(\theta_1 - \theta_2)\big\rangle \\
\psi'(t) &= R\big\langle f'\big(\theta_2 + t(\theta_1 - \theta_2)\big) - f'(\theta_2), \theta_1 - \theta_2\big\rangle \\
\psi''(t) &= R\big\langle \theta_2 - \theta_1, f''\big(\theta_2 + t(\theta_1 - \theta_2)\big)(\theta_1 - \theta_2)\big\rangle,
\end{aligned}
$$

Thus $\varphi'(0) = \psi'(0) = 0$ and $\varphi''(t) \leqslant \psi''(t)$, leading to $\varphi(1) \leqslant \psi(1)$ 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: $\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) = \frac{1}{n}\sum_{k=1}^n \big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big] + \frac{1}{\gamma n}(\theta_0 - \theta_*) + \frac{1}{\gamma n}(\theta_* - \theta_n) = O(1/\sqrt{n})$

- Step 3: $\Big\| f'\big(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\big) - \frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) \Big\|_2$
$= O\big(f(\bar{\theta}_n) - f(\theta_*)\big) = O(1/\sqrt{n})$ using self-concordance

- Step 4a: if f μ-strongly convex, $f(\bar{\theta}_n) - f(\theta_*) \leqslant \frac{1}{2\mu}\big\| f'(\bar{\theta}_n)\big\|_2^2$

- 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 $|\theta^\top \Phi(x_n)| \leqslant M$ (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** $1/\big(2R^2\sqrt{n}\big)$
  - with $R =$ radius of data (Bach, 2013):

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

  - 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 $|\theta^\top \Phi(x_n)| \leqslant M$ (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** $1/\big(2R^2\sqrt{n}\big)$
  - with $R =$ radius of data (Bach, 2013):

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

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

Least-mean-square algorithm

- **Least-squares**: $f(\theta) = \frac{1}{2}\mathbb{E}\big[(y_n - \langle \Phi(x_n), \theta\rangle)^2\big]$ 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 $\mathbb{E}\big[\Phi(x_n) \otimes \Phi(x_n)\big] = H \succcurlyeq \mu \cdot \mathrm{Id}$

Least-mean-square algorithm

- **Least-squares**: $f(\theta) = \frac{1}{2}\mathbb{E}\big[(y_n - \langle \Phi(x_n), \theta\rangle)^2\big]$ 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 $\mathbb{E}\big[\Phi(x_n) \otimes \Phi(x_n)\big] = H \succcurlyeq \mu \cdot \mathrm{Id}$

- **New analysis for averaging and constant step-size** $\gamma = 1/(4R^2)$
  - Assume $\|\Phi(x_n)\| \leqslant R$ and $|y_n - \langle \Phi(x_n), \theta_*\rangle| \leqslant \sigma$ almost surely
  - No assumption regarding lowest eigenvalues of H
  - Main result: $\boxed{\mathbb{E} f(\bar{\theta}_{n-1}) - f(\theta_*) \leqslant \frac{4\sigma^2 d}{n} + \frac{4R^2\|\theta_0 - \theta_*\|^2}{n}}$

- **Matches statistical lower bound** (Tsybakov, 2003)
  - Non-asymptotic robust version of Györfi and Walk (1996)

Least-squares - Proof technique - I

- LMS recursion:

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

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

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

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

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

- Infinite expansion of Aguech, Moulines, and Priouret (2000) in powers of γ

Least-squares - Proof technique - II

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

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

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

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

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

Least-squares - Proof technique - III

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

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

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

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

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

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

- **Variance** (next slide)

Least-squares - Proof technique - III

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

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

- **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 γ
  - LMS recursion:

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

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

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

- Expansion of the difference:

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

Least-squares - Proof technique - IV

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

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

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

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

- Expansion of the difference:

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

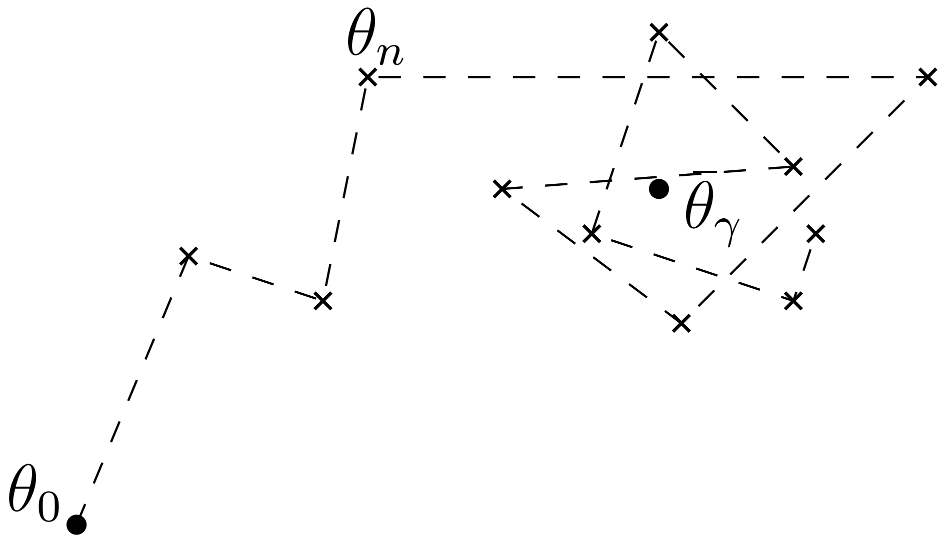
  - New noise process
  - May continue the expansion infinitely many times

Markov chain interpretation of constant step sizes

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

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

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



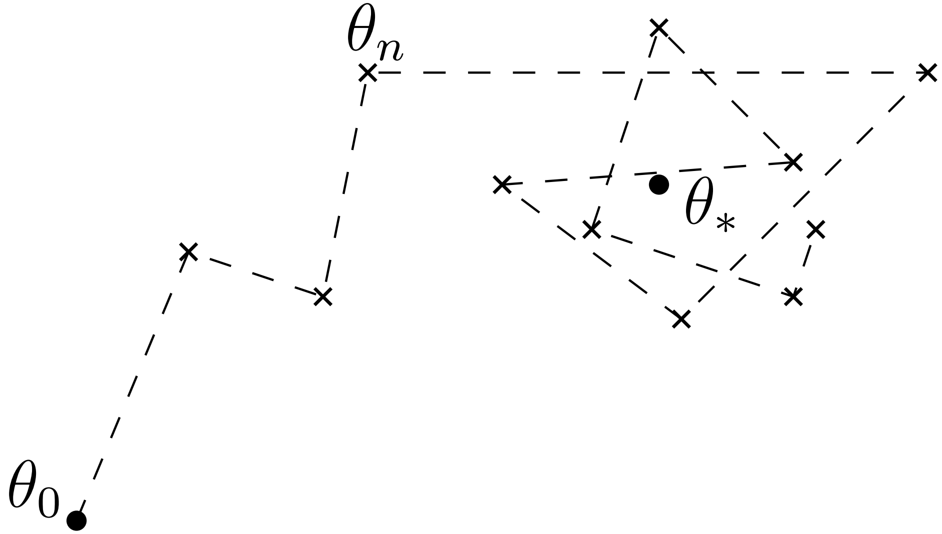
Markov chain interpretation of constant step sizes

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

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

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

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



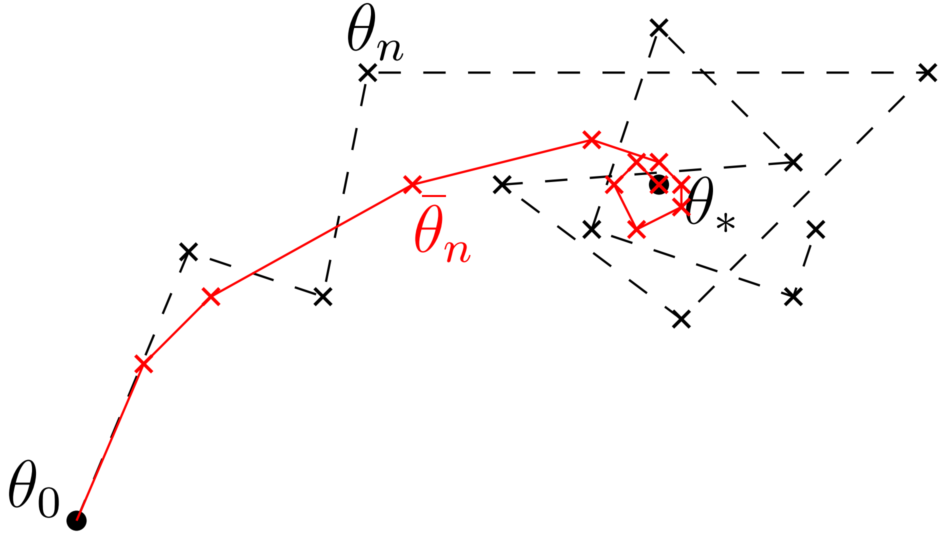
Markov chain interpretation of constant step sizes

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

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

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

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



Markov chain interpretation of constant step sizes

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

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

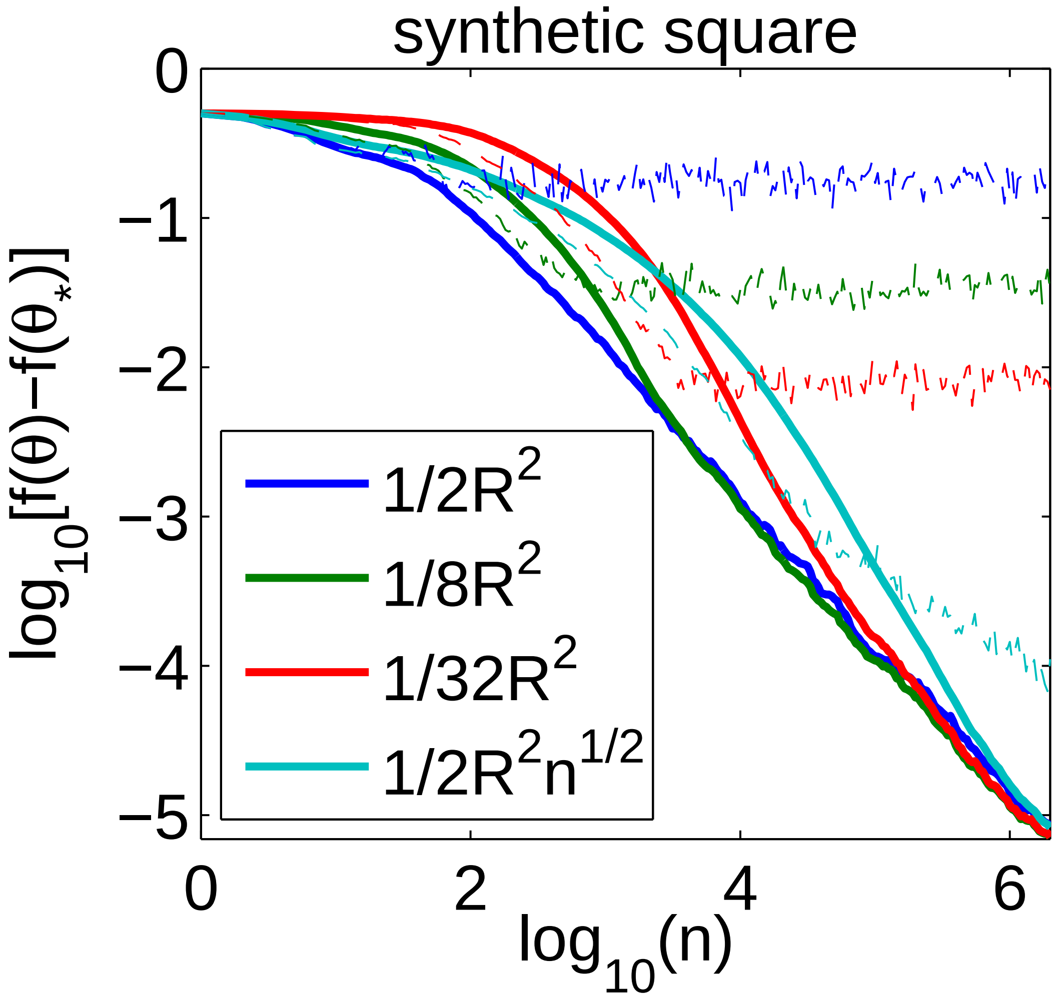
- The sequence $(\theta_n)_n$ is a homogeneous Markov chain
  - convergence to a stationary distribution π_γ
  - with expectation $\bar{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$

- **For least-squares,** $\bar{\theta}_\gamma = \theta_*$
  - θ_n does not converge to θ_* but oscillates around it
  - oscillations of order $\sqrt{\gamma}$

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

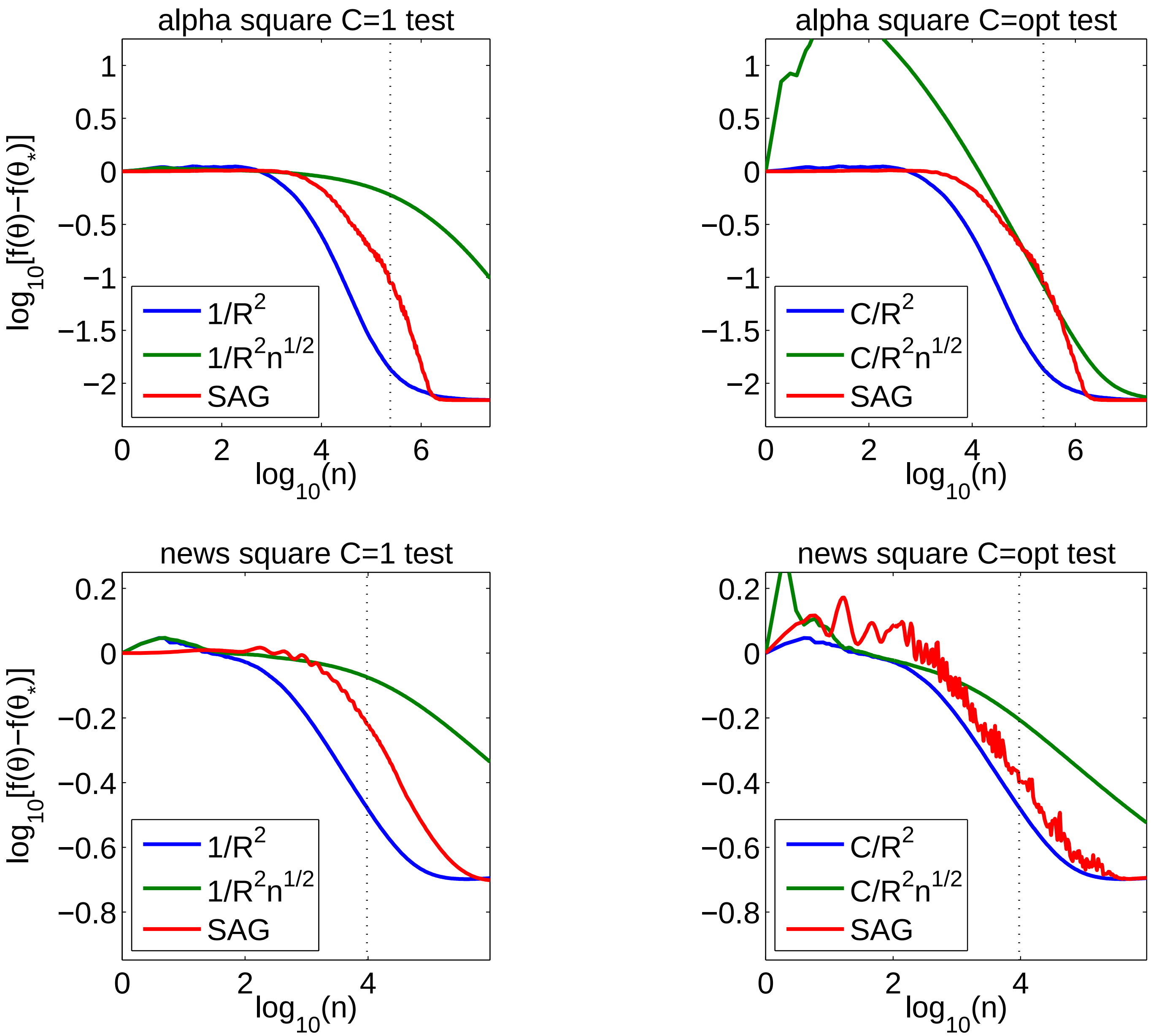
Simulations - synthetic examples

- Gaussian distributions - $d = 20$



Simulations - benchmarks

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



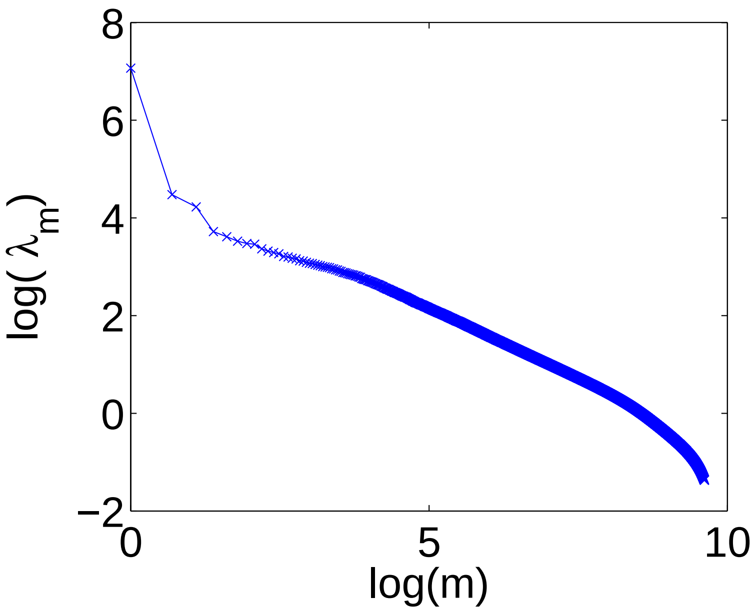
Optimal bounds for least-squares?

- **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

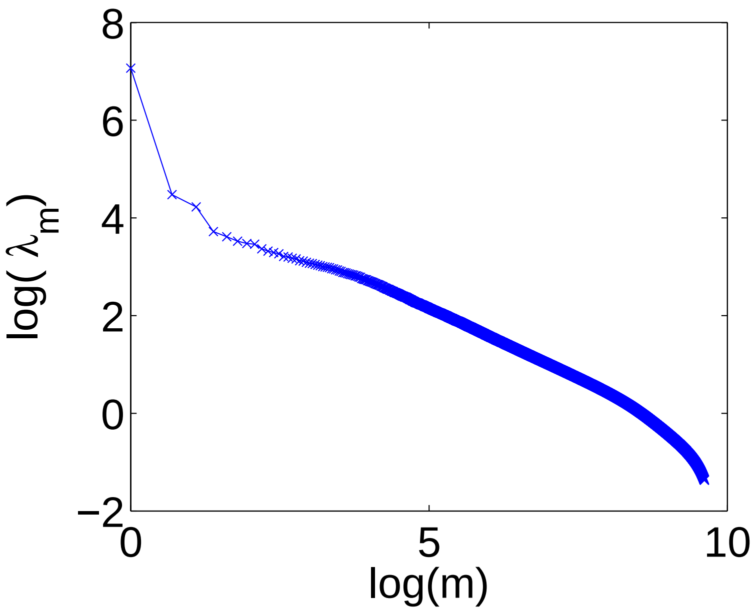
Finer assumptions (Dieuleveut and Bach, 2014)

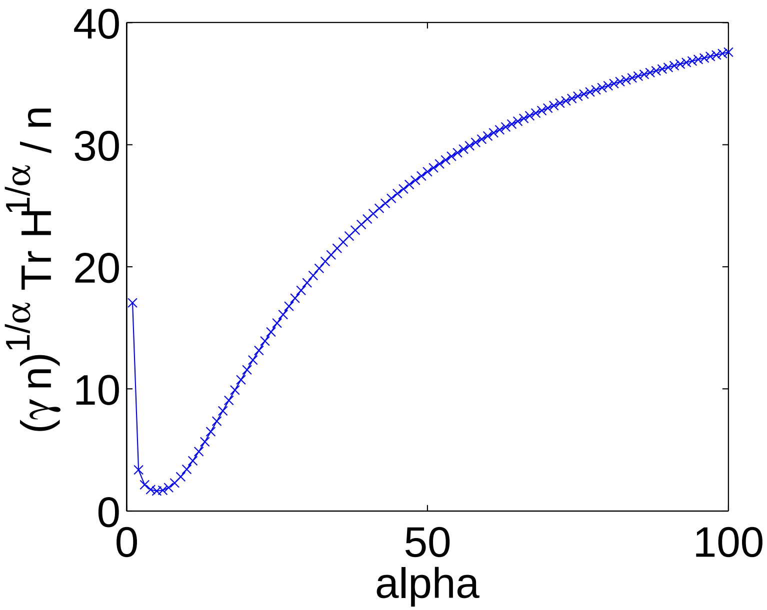
- **Covariance eigenvalues**
  - Pessimistic assumption: all eigenvalues λ_m less than a constant
  - Actual decay as $\lambda_m = o(m^{-\alpha})$ with $\operatorname{tr} H^{1/\alpha} = \sum_m \lambda_m^{1/\alpha}$ small



Finer assumptions (Dieuleveut and Bach, 2014)

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





Finer assumptions (Dieuleveut and Bach, 2014)

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

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

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

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

  - Previous results: $\alpha = +\infty$ and $r = 1/2$
  - Valid for all α 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**: $\Phi(x), \theta \in \mathcal{H}$

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

- **If $\theta_0 = 0$, θ_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**: $\Phi(x), \theta \in \mathcal{H}$

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

- **If $\theta_0 = 0$, θ_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 $k(x,x') = \sum_j \varphi_j^2 \cos\big[2j\pi(x-x')\big]$
  - Optimal prediction function θ_* 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 $k(x, x') = \sum_j \varphi_j^2 \cos\big[2j\pi(x - x')\big]$
  - Optimal prediction function θ_* has norm $\|\theta_*\|^2 = \sum_j |\mathcal{F}(\theta_*)_j|^2 \varphi_j^{-2}$
  - Depending on smoothness, may or may not be finite

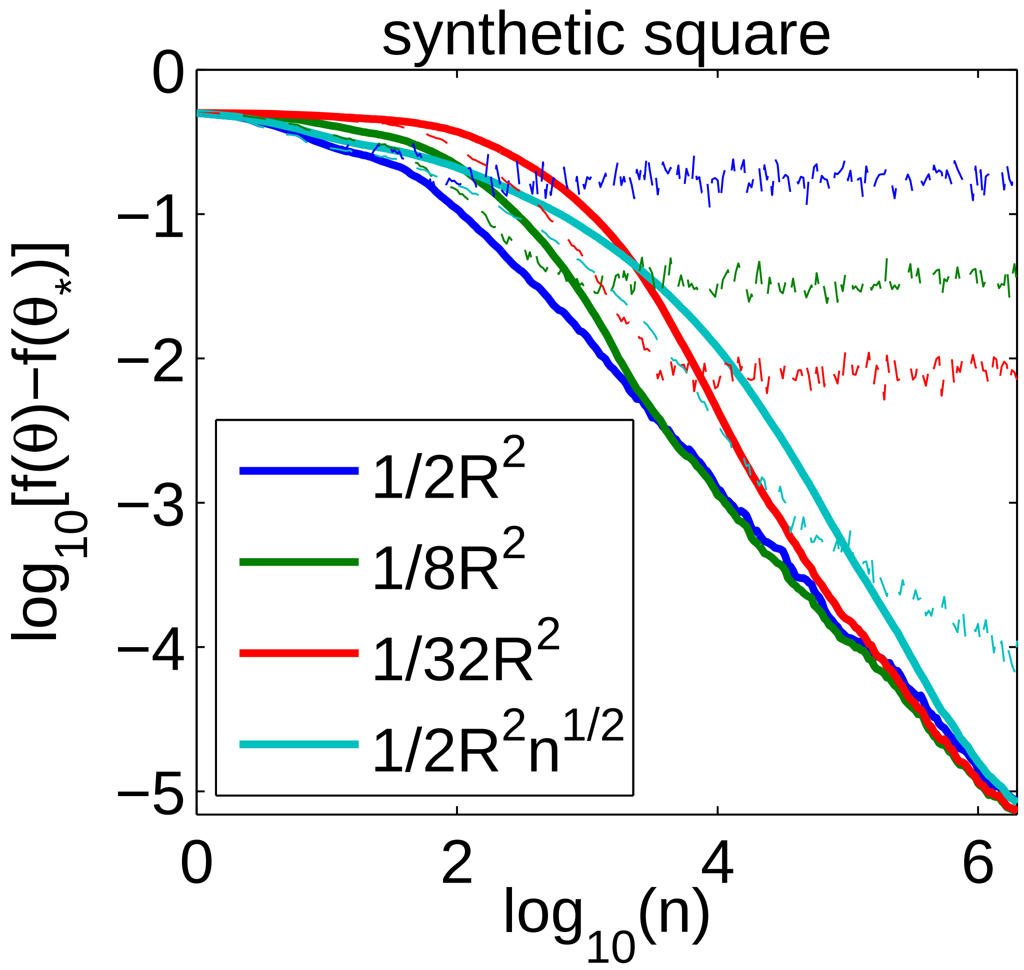
- Adapted norm $\|H^{1/2-r}\theta_*\|^2 = \sum_j |\mathcal{F}(\theta_*)_j|^2 \varphi_j^{-4r}$ may be finite

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

- **Same effect than ℓ_2-regularization with weight λ equal to** $\dfrac{1}{\gamma n}$

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 $n \to \infty$ and $\gamma \to 0$

- **Variance** (e.g., starting from the solution)

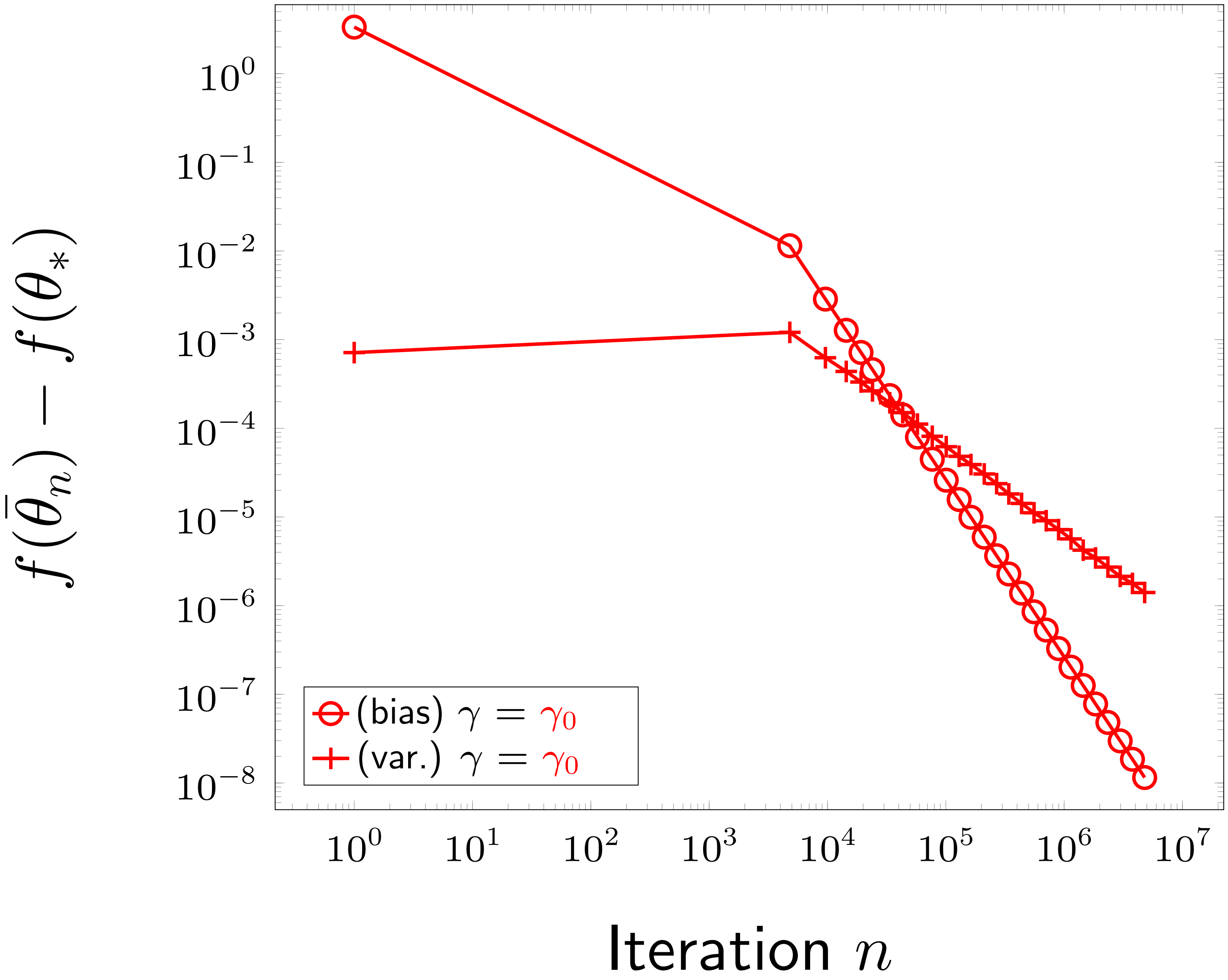
$$f(\bar{\theta}_n) - f(\theta_*) \sim \frac{1}{n} \mathbb{E}\Big[\varepsilon^2 \, \Phi(x)^\top H^{-1} \Phi(x)\Big]$$

  – NB: if noise ε is independent, then we obtain $\frac{d\sigma^2}{n}$
  – Exponentially decaying remainder terms (strongly convex problems)

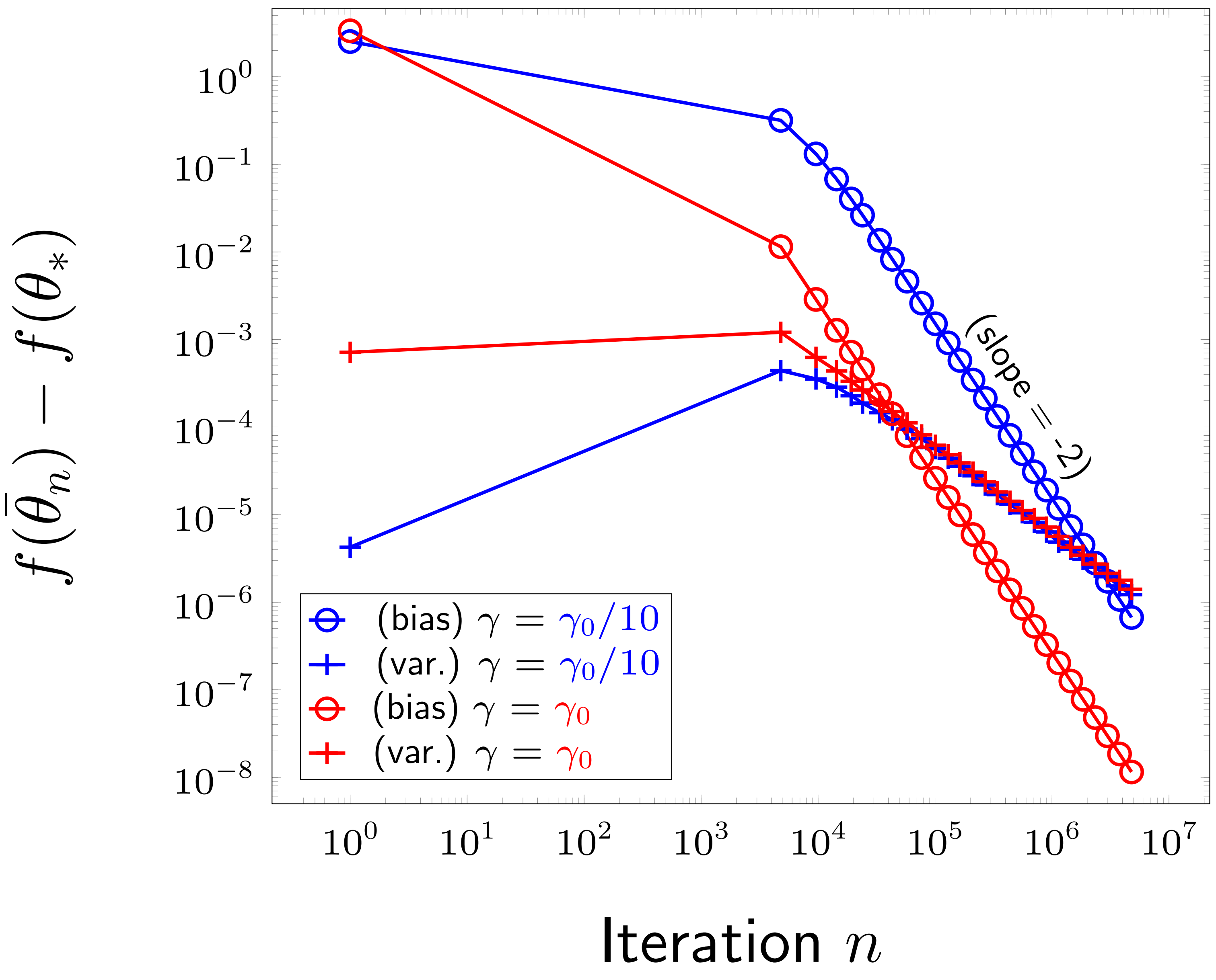
- **Bias** (e.g., no noise)

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

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

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

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

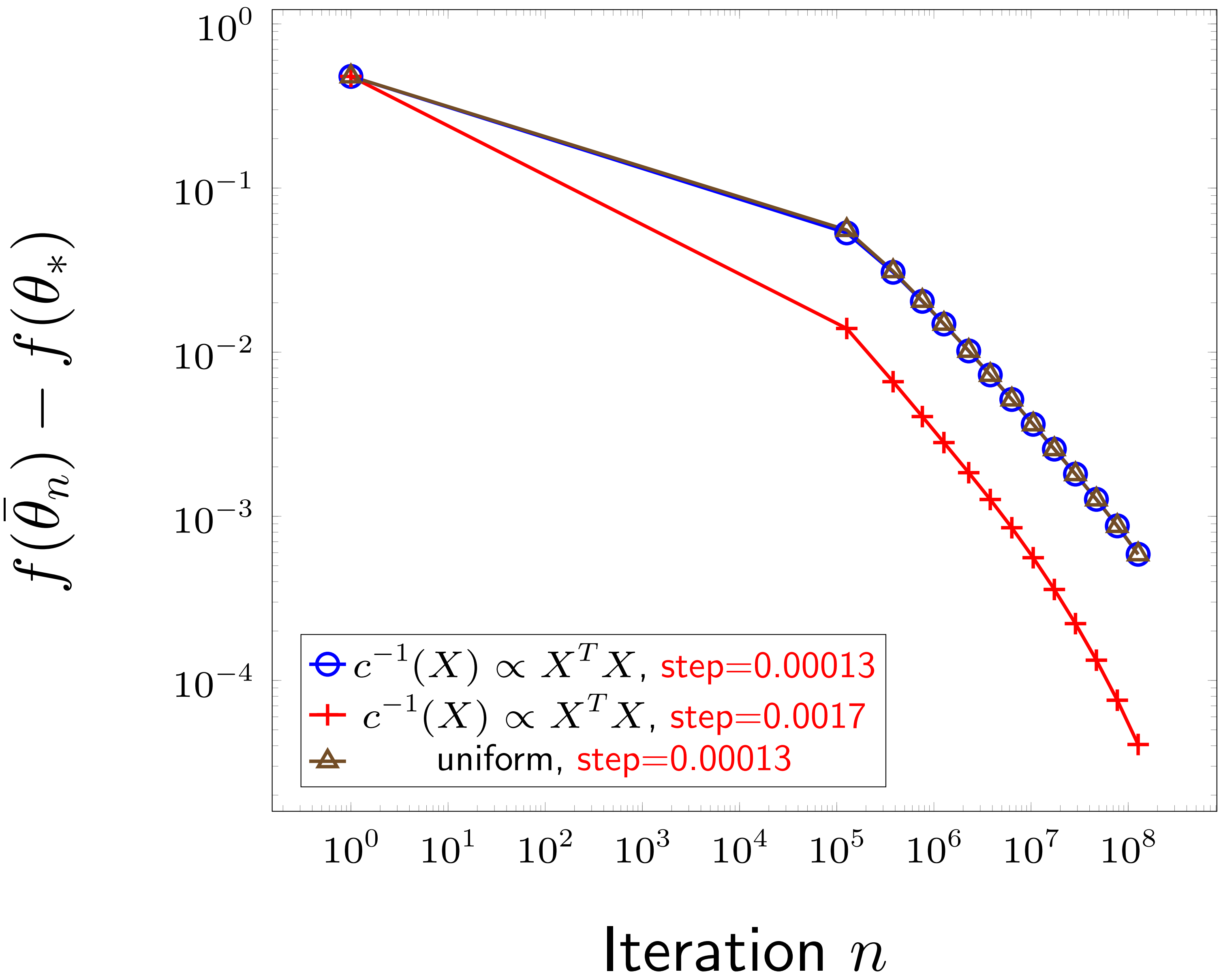
$$\mathbb{E}_{p(x)p(y|x)}|y - \Phi(x)^\top\theta|^2 = \mathbb{E}_{q(x)p(y|x)}\frac{dp(x)}{dq(x)}|y - \Phi(x)^\top\theta|^2$$

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

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

  – Simpy allows biggest possible step size $\gamma < \frac{2}{\operatorname{tr} H}$
  – 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 θ_*) $= \dfrac{\sigma^2 d}{n}$
  - **Bias** (no noise) $= \min\left\{\dfrac{R^2\|\theta_0 - \theta_*\|^2}{n}, \dfrac{R^4\langle\theta_0 - \theta_*, H^{-1}(\theta_0 - \theta_*)\rangle}{n^2}\right\}$

Achieving optimal bias and variance terms

- **Current results with averaged SGD** (ill-conditioned problems)
  - **Variance** (starting from optimal θ_*) $= \dfrac{\sigma^2 d}{n}$
  - **Bias** (no noise) $= \dfrac{R^2\|\theta_0 - \theta_*\|^2}{n}$

Achieving optimal bias and variance terms

- **Current results with averaged SGD** (ill-conditioned problems)
  - **Variance** (starting from optimal θ_*) $= \dfrac{\sigma^2 d}{n}$
  - **Bias** (no noise) $= \dfrac{R^2\|\theta_0 - \theta_*\|^2}{n}$

	Bias	Variance
Averaged gradient descent (Bach and Moulines, 2013)	$\dfrac{R^2\|\theta_0 - \theta_*\|^2}{n}$	$\dfrac{\sigma^2 d}{n}$

Achieving optimal bias and variance terms

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

- **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
Averaged gradient descent (Bach and Moulines, 2013)	$\frac{R^2\\|\theta_0-\theta_*\\|^2}{n}$	$\frac{\sigma^2 d}{n}$
Accelerated gradient descent (Nesterov, 1983)	$\frac{R^2\\|\theta_0-\theta_*\\|^2}{n^2}$	$\sigma^2 d$
"Between" averaging and acceleration (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
Averaged gradient descent (Bach and Moulines, 2013)	$\frac{R^2\|\theta_0-\theta_*\|^2}{n}$	$\frac{\sigma^2 d}{n}$
Accelerated gradient descent (Nesterov, 1983)	$\frac{R^2\|\theta_0-\theta_*\|^2}{n^2}$	$\sigma^2 d$
"Between" averaging and acceleration (Flammarion and Bach, 2015)	$\frac{R^2\|\theta_0-\theta_*\|^2}{n^{1+\alpha}}$	$\frac{\sigma^2 d}{n^{1-\alpha}}$
Averaging and acceleration (Dieuleveut, Flammarion, and Bach, 2016)	$\frac{R^2\|\theta_0-\theta_*\|^2}{n^2}$	$\frac{\sigma^2 d}{n}$

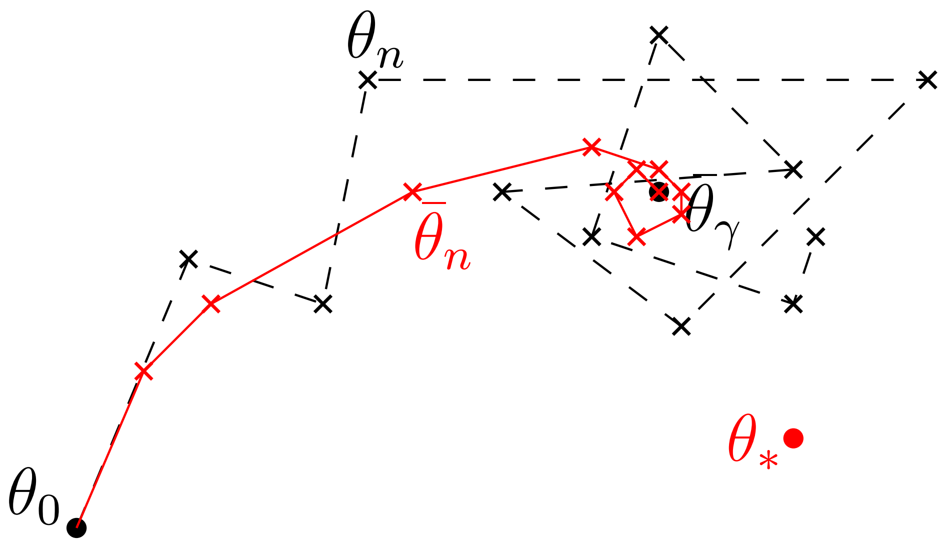
Beyond least-squares - Markov chain interpretation

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

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 π_γ such that $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$
  - When f' is not linear, $f'(\int \theta \pi_\gamma(\mathrm{d}\theta)) \neq \int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$

- **θ_n oscillates around the wrong value $\bar{\theta}_\gamma \neq \theta_*$**



Beyond least-squares - Markov chain interpretation

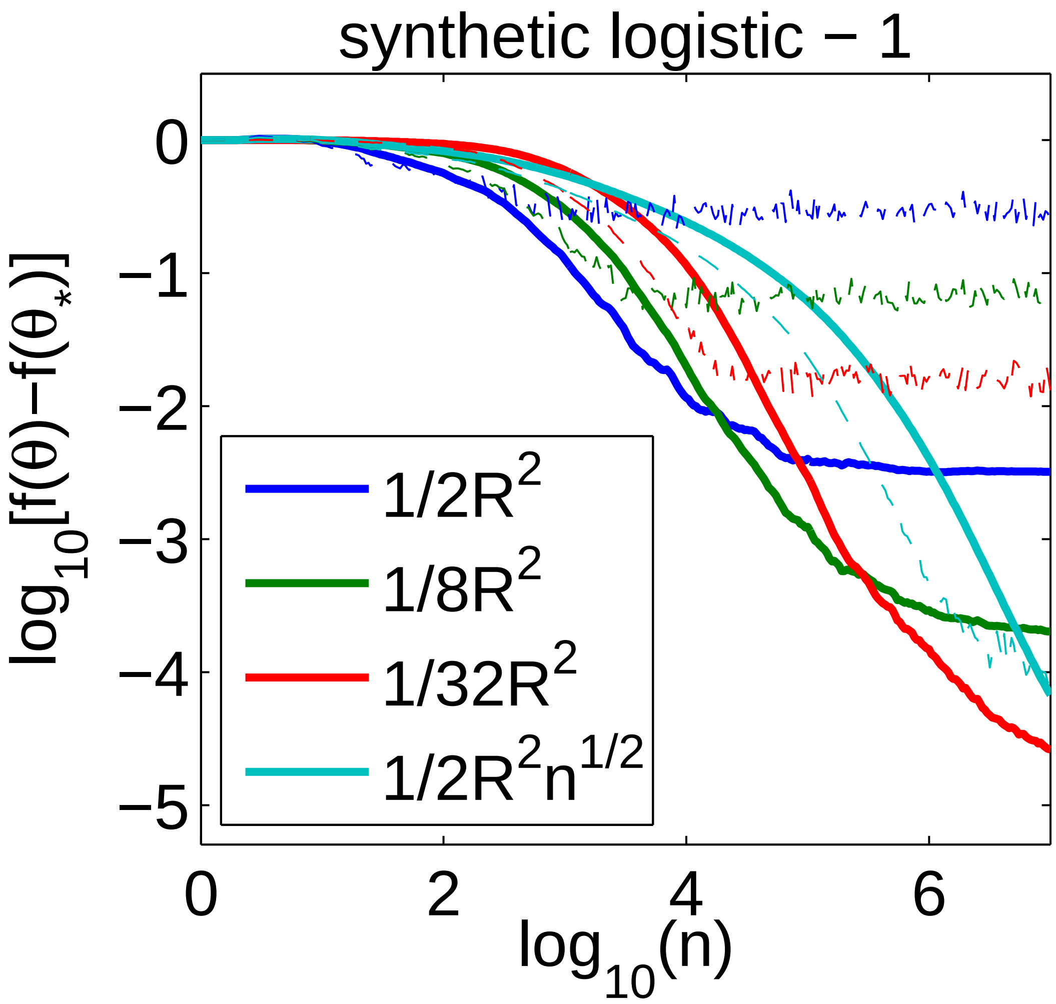
- Recursion $\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1})$ also defines a Markov chain
  - Stationary distribution π_γ such that $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$
  - When f' is not linear, $f'(\int \theta \pi_\gamma(\mathrm{d}\theta)) \neq \int f'(\theta)\pi_\gamma(\mathrm{d}\theta)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

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

$$
\begin{aligned}
g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\
&= f(\tilde{\theta}) + \langle \mathbb{E} f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, \mathbb{E} f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\
&= \mathbb{E}\Big[f(\tilde{\theta}) + \langle f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle\Big]
\end{aligned}
$$

Restoring convergence through online Newton steps

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

$$
\begin{aligned}
g(\theta) &= f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\
&= f(\tilde{\theta}) + \langle \mathbb{E} f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, \mathbb{E} f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle \\
&= \mathbb{E}\Big[f(\tilde{\theta}) + \langle f_n'(\tilde{\theta}), \theta - \tilde{\theta}\rangle + \tfrac{1}{2}\langle \theta - \tilde{\theta}, f_n''(\tilde{\theta})(\theta - \tilde{\theta})\rangle\Big]
\end{aligned}
$$

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

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

  – $f_n''(\tilde{\theta}) = \ell''(y_n, \langle \tilde{\theta}, \Phi(x_n)\rangle)\Phi(x_n) \otimes \Phi(x_n)$ 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 $\varphi : u \mapsto \log(1 + e^{-u})$:

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

  – NB: difference with regular self-concordance: $|\varphi'''(u)| \leqslant 2\varphi''(u)^{3/2}$

- Using novel high-probability convergence results for regular averaged stochastic gradient descent

- Requires assumption on the kurtosis in every direction, i.e.,

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

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

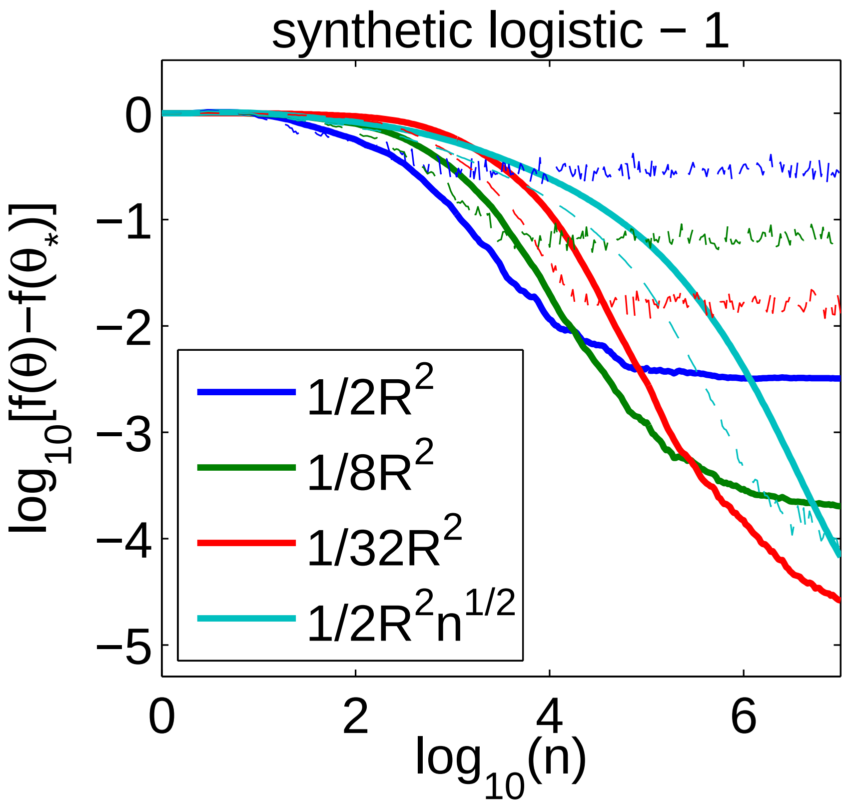
$$\begin{cases} \theta_n = \theta_{n-1} - \gamma\big[f_n'(\bar\theta_{n-1}) + f_n''(\bar\theta_{n-1})(\theta_{n-1} - \bar\theta_{n-1})\big] \\ \bar\theta_n = \bar\theta_{n-1} + \frac{1}{n}(\theta_n - \bar\theta_{n-1}) \end{cases}$$

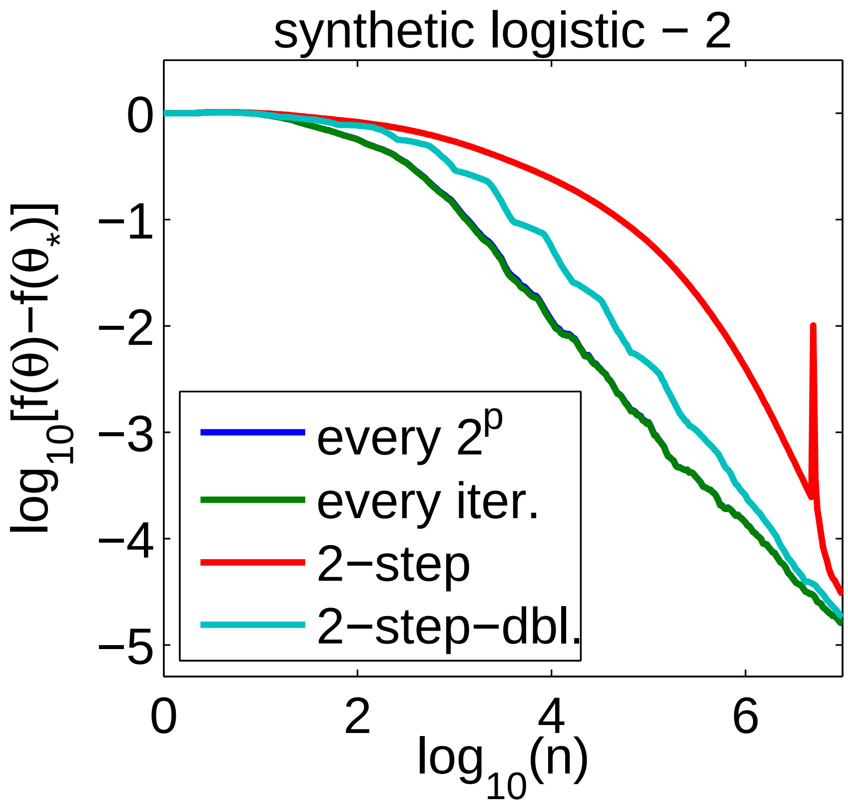
- Instance of two-time-scale stochastic approximation (Borkar, 1997)
  - Given $\bar\theta$, $\theta_n = \theta_{n-1} - \gamma\big[f_n'(\bar\theta) + f_n''(\bar\theta)(\theta_{n-1} - \bar\theta)\big]$ defines a homogeneous Markov chain (fast dynamics)
  - $\bar\theta_n$ is updated at rate $1/n$ (slow dynamics)

- **Difficulty**: preserving robustness to ill-conditioning

Simulations - synthetic examples

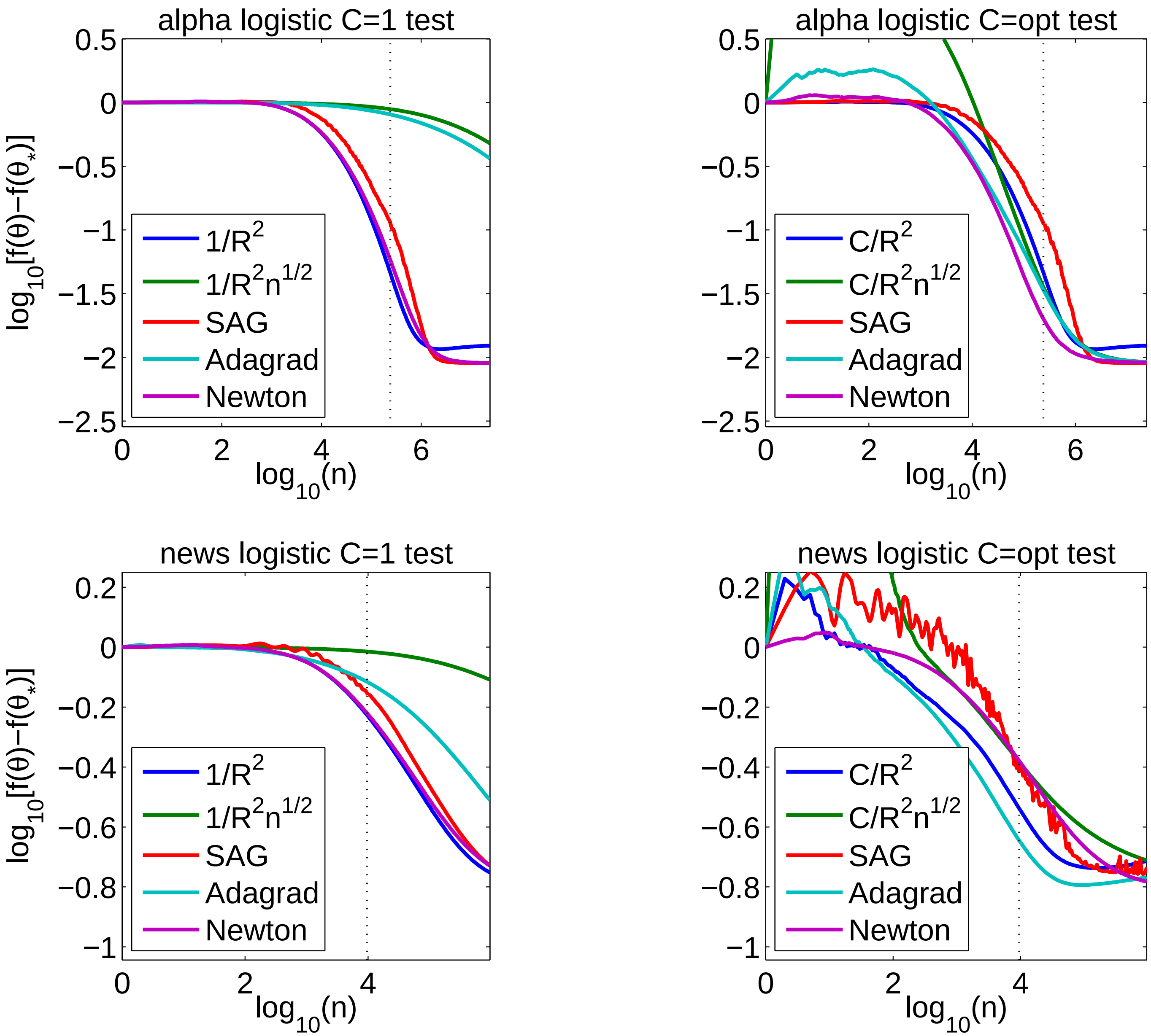
- Gaussian distributions - $d = 20$





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
  - μ 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
  - μ 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. **Classical stochastic approximation**

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

Outline - II

4. **Non-smooth stochastic approximation**
   - Stochastic (sub)gradient and averaging
   - Non-asymptotic results and lower bounds
   - Strongly convex vs. non-strongly convex

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 ℓ_2-norm) to avoid overfitting
- **Goal**: minimize $g(\theta) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n f_i(\theta)$

Stochastic vs. deterministic methods

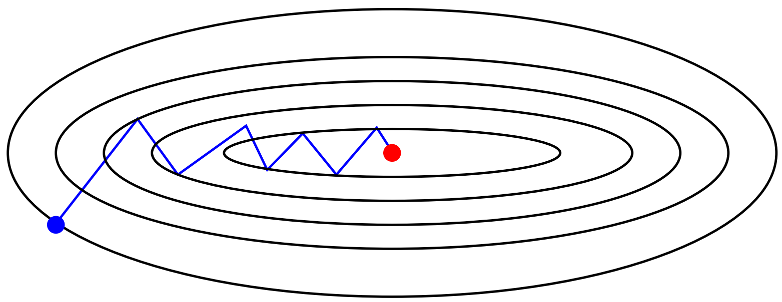
- Minimizing $g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ with $f_i(\theta) = \ell\big(y_i, \theta^\top \Phi(x_i)\big) + \mu\Omega(\theta)$

- 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 $O(e^{-\alpha t})$
  - Iteration complexity is linear in n *(with line search)*

Stochastic vs. deterministic methods

- Minimizing $g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ with $f_i(\theta) = \ell\big(y_i, \theta^\top \Phi(x_i)\big) + \mu\Omega(\theta)$

- 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 $f_i(\theta) = \ell\big(y_i, \theta^\top \Phi(x_i)\big) + \mu\Omega(\theta)$

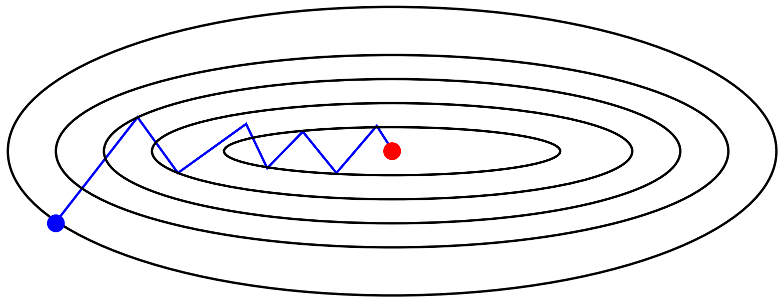
- Batch gradient descent: $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n}\sum_{i=1}^{n} f_i'(\theta_{t-1})$
  - Linear (e.g., exponential) convergence rate in $O(e^{-\alpha t})$
  - Iteration complexity is linear in n *(with line search)*

- 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 $O(1/t)$
  - Iteration complexity is independent of n *(step size selection?)*

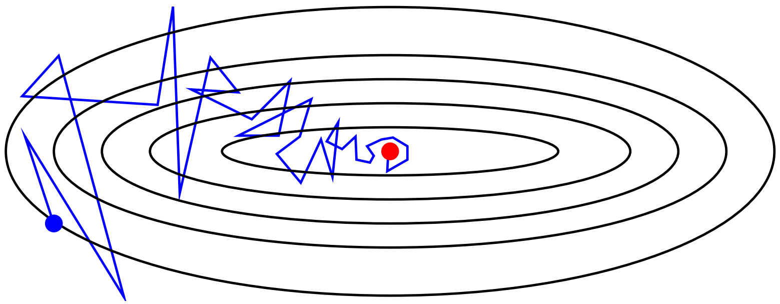
Stochastic vs. deterministic methods

- Minimizing $g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ with $f_i(\theta) = \ell\big(y_i, \theta^\top \Phi(x_i)\big) + \mu\Omega(\theta)$

- Batch gradient descent: $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n}\sum_{i=1}^{n} f_i'(\theta_{t-1})$

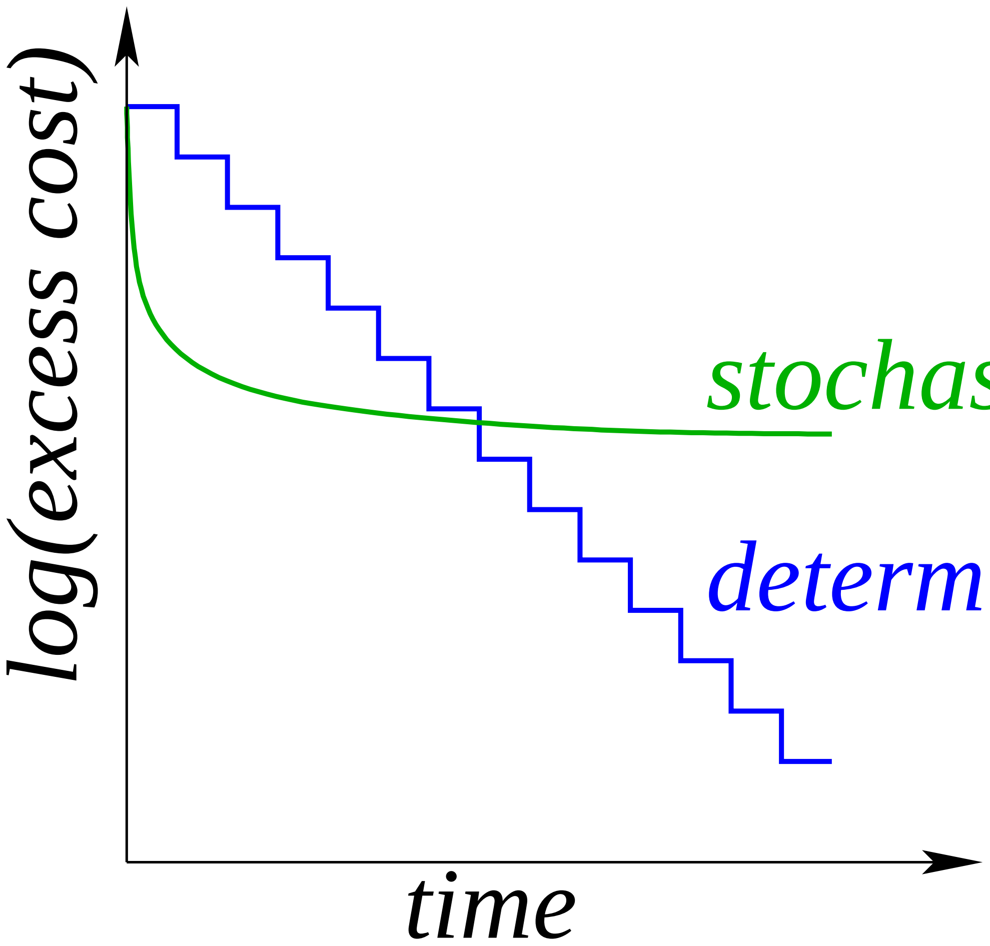


- Stochastic gradient descent: $\theta_t = \theta_{t-1} - \gamma_t f_{i(t)}'(\theta_{t-1})$



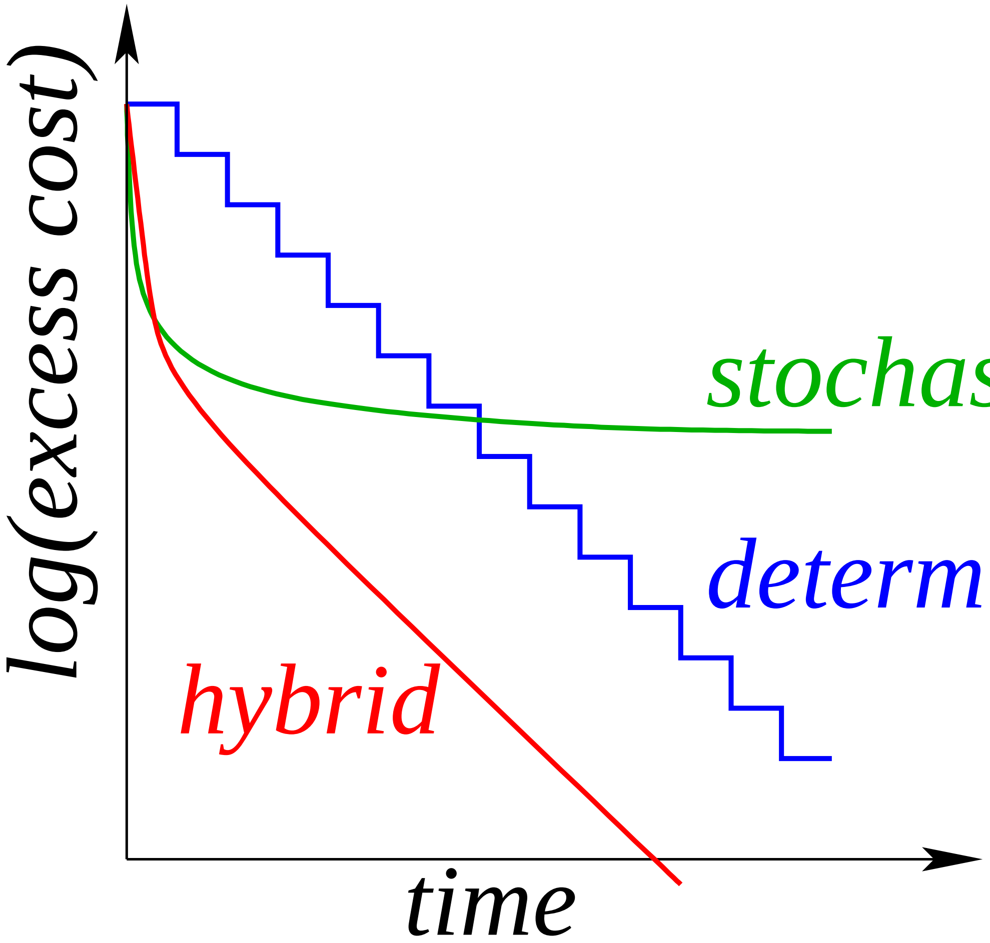
Stochastic vs. deterministic methods

- **Goal** = **best of both worlds**: Linear rate with $O(1)$ iteration cost
  Robustness to step size



Stochastic vs. deterministic methods

- **Goal** = **best of both worlds**: Linear rate with $O(1)$ iteration cost
  Robustness to step size



Accelerating gradient methods - Related work

- **Nesterov acceleration**
  - Nesterov (1983, 2004)
  - Better linear rate but still $O(n)$ iteration cost

- **Hybrid methods, incremental average gradient, increasing batch size**
  - Bertsekas (1997); Blatt et al. (2008); Friedlander and Schmidt (2011)
  - Linear rate, but iterations make full passes through the data.

Accelerating gradient methods - Related work

- **Momentum, gradient/iterate averaging, stochastic version of accelerated batch gradient methods**
  - Polyak and Juditsky (1992); Tseng (1998); Sunehag et al. (2009); Ghadimi and Lan (2010); Xiao (2010)
  - Can improve constants, but still have sublinear $O(1/t)$ rate

- **Constant step-size stochastic gradient (SG), accelerated SG**
  - Kesten (1958); Delyon and Juditsky (1993); Solodov (1998); Nedic and Bertsekas (2000)
  - Linear convergence, but only up to a fixed tolerance.

- **Stochastic methods in the dual**
  - Shalev-Shwartz and Zhang (2012)
  - Similar linear rate but limited choice for the f_i's

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 $y_i^t = \begin{cases} f_i'(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$

- 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

Stochastic average gradient - Convergence analysis

- **Assumptions**
  - Each f_i is R^2-smooth, $i = 1, \dots, n$
  - $g = \frac{1}{n}\sum_{i=1}^{n} f_i$ is μ-strongly convex (with potentially $\mu = 0$)
  - constant step size $\gamma_t = 1/(16R^2)$
  - initialization with one pass of averaged SGD

Stochastic average gradient - Convergence analysis

- **Assumptions**
  - Each f_i is R^2-smooth, $i = 1, \dots, n$
  - $g = \frac{1}{n} \sum_{i=1}^n f_i$ is μ-strongly convex (with potentially $\mu = 0$)
  - constant step size $\gamma_t = 1/(16R^2)$
  - initialization with one pass of averaged SGD

- **Strongly convex case** (Le Roux et al., 2012, 2013)

$$\mathbb{E}\big[g(\theta_t) - g(\theta_*)\big] \leqslant \Big(\frac{8\sigma^2}{n\mu} + \frac{4R^2 \|\theta_0 - \theta_*\|^2}{n} \Big) \exp\Big(- t \min\Big\{ \frac{1}{8n}, \frac{\mu}{16R^2} \Big\} \Big)$$

  - Linear (exponential) convergence rate with $O(1)$ iteration cost
  - After one pass, reduction of cost by $\exp\Big(- \min\Big\{ \frac{1}{8}, \frac{n\mu}{16R^2} \Big\} \Big)$

Stochastic average gradient - Convergence analysis

- **Assumptions**
  - Each f_i is R^2-smooth, $i = 1, \dots, n$
  - $g = \frac{1}{n}\sum_{i=1}^n f_i$ is μ-strongly convex (with potentially $\mu = 0$)
  - constant step size $\gamma_t = 1/(16R^2)$
  - initialization with one pass of averaged SGD

- **Non-strongly convex case** (Le Roux et al., 2013)

$$\mathbb{E}\big[g(\theta_t) - g(\theta_*)\big] \leqslant 48 \frac{\sigma^2 + R^2\|\theta_0 - \theta_*\|^2}{\sqrt{n}} \frac{n}{t}$$

  - Improvement over regular batch and stochastic gradient
  - Adaptivity to potentially hidden strong convexity

Convergence analysis - Proof sketch

- **Main step**: find "good" Lyapunov function $J(\theta_t, y_1^t, \dots, y_n^t)$
  - such that $\mathbb{E}\big[J(\theta_t, y_1^t, \dots, y_n^t) | \mathcal{F}_{t-1}\big] < J(\theta_{t-1}, y_1^{t-1}, \dots, y_n^{t-1})$
  - 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, μ, L)
  - Check validity with symbolic computations

Rate of convergence comparison

- Assume that $L = 100$, $\mu = .01$, and $n = 80000$
  - Full gradient method has rate
$$\left(1 - \frac{\mu}{L}\right) = 0.9999$$
  - Accelerated gradient method has rate
$$\left(1 - \sqrt{\frac{\mu}{L}}\right) = 0.9900$$
  - Running n iterations of SAG for the same cost has rate
$$\left(1 - \frac{1}{8n}\right)^n = 0.8825$$
  - *Fastest possible* first-order method has rate
$$\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 = 0.9608$$
- **Beating two lower bounds** (with additional assumptions)
  - (1) stochastic gradient and (2) full gradient

Stochastic average gradient
Implementation details and extensions

- The algorithm can use sparsity in the features to reduce the storage and iteration cost

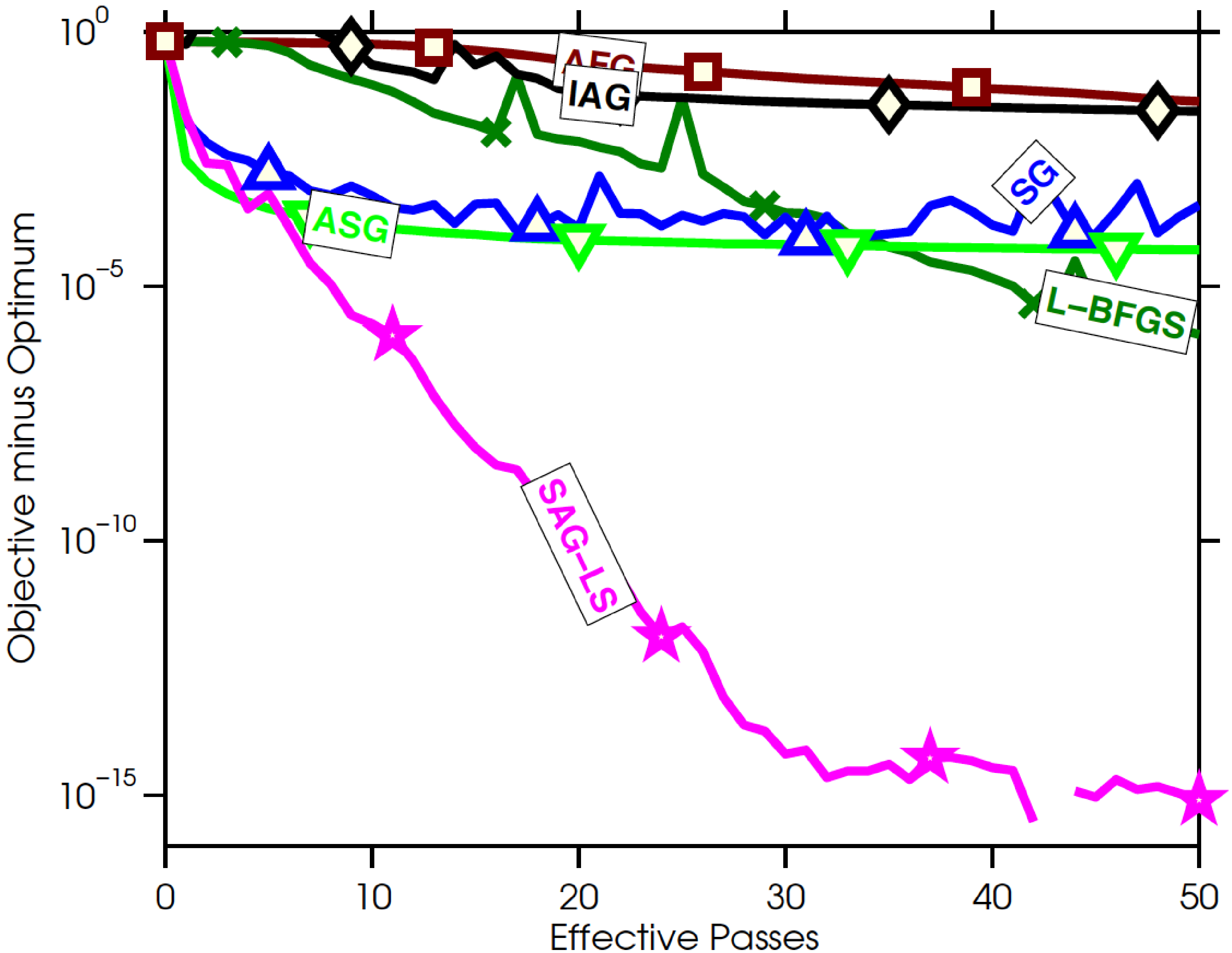
- Grouping functions together can further reduce the memory requirement

- We have obtained good performance when R^2 is not known with a heuristic line-search

- Algorithm allows non-uniform sampling

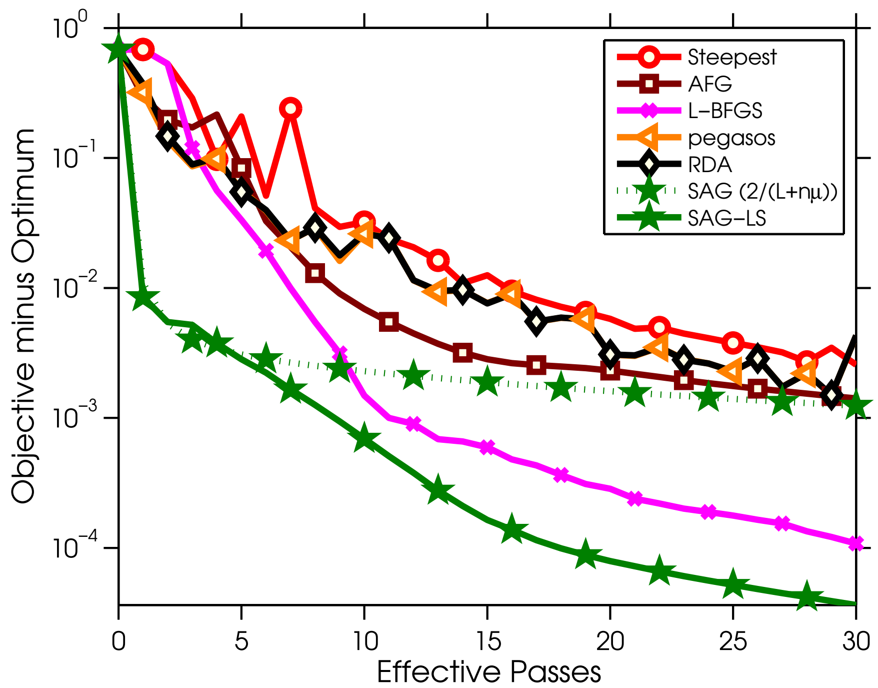
- Possibility of making proximal, coordinate-wise, and Newton-like variants

spam dataset (n = 92 189, d = 823 470)

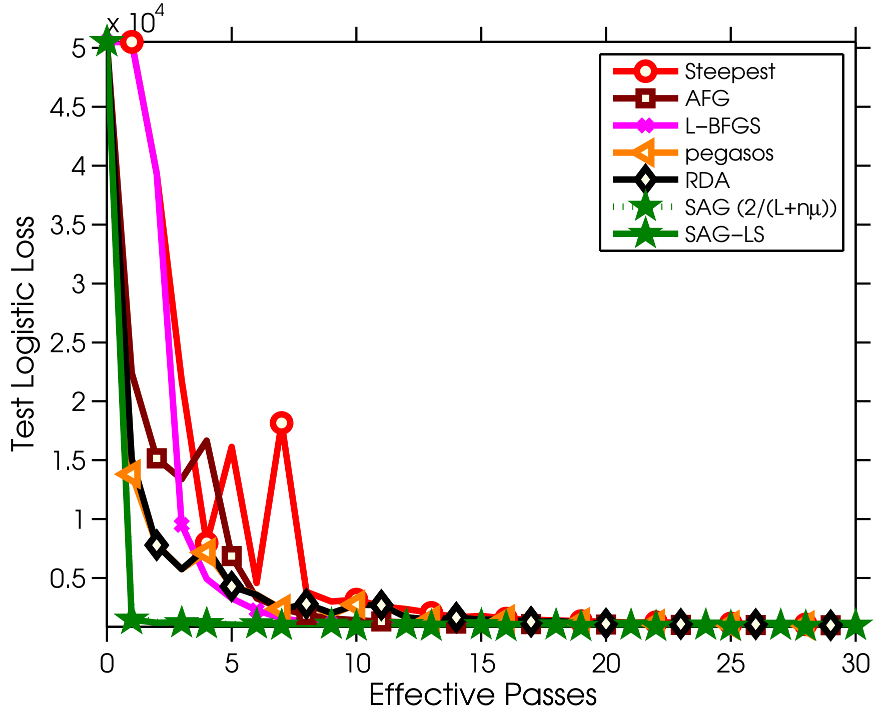


`protein` dataset ($n = 145751$, $d = 74$)

- Dataset split in two (training/testing)



Training cost



Testing cost

Extensions and related work

- **Exponential convergence rate for strongly convex problems**
- **Need to store gradients**
  - SVRG (Johnson and Zhang, 2013)
- **Adaptivity to non-strong convexity**
  - SAGA (Defazio, Bach, and Lacoste-Julien, 2014)
- **Simple proof**
  - SVRG, SAGA
- **Lower bounds**
  - Agarwal and Bottou (2014)

Variance reduction

- **Principle**: reducing variance of sample of X by using a sample from another random variable Y with known expectation

$$Z_\alpha = \alpha(X - Y) + \mathbb{E}Y$$

  - $\mathbb{E}Z_\alpha = \alpha\mathbb{E}X + (1-\alpha)\mathbb{E}Y$
  - $\operatorname{var} Z_\alpha = \alpha^2\big[\operatorname{var} X + \operatorname{var} Y - 2\operatorname{cov}(X,Y)\big]$
  - $\alpha = 1$: 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

$$Z_\alpha = \alpha(X - Y) + \mathbb{E}Y$$

  - $\mathbb{E}Z_\alpha = \alpha\mathbb{E}X + (1-\alpha)\mathbb{E}Y$
  - $\mathrm{var}\, Z_\alpha = \alpha^2\big[\,\mathrm{var}\, X + \mathrm{var}\, Y - 2\mathrm{cov}(X,Y)\big]$
  - $\alpha = 1$: no bias, $\alpha < 1$: potential bias (but reduced variance)
  - Useful if Y positively correlated with X

- **Application to gradient estimation : SVRG** (Johnson and Zhang, 2013)

  - Estimating the averaged gradient $g'(\theta) = \frac{1}{n}\sum_{i=1}^n f_i'(\theta)$
  - Using the gradients of a previous iterate $\tilde{\theta}$

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: $\theta_t = \theta_{t-1} - \gamma \Big[f'_{i_t}(\theta_{t-1}) - f'_{i_t}(\tilde{\theta}) + g'(\tilde{\theta}) \Big]$

- **Proposition**: If each f_i is R^2-smooth and $g = \frac{1}{n}\sum_{i=1}^n$ is μ-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**: $\mathbb{E}\| f_i'(\theta) - f_i'(\theta_*)\|^2 \leqslant 2R^2\big[g(\theta) - g(\theta_*)\big]$
  - Proof: $\mathbb{E}\| f_i'(\theta) - f_i'(\theta_*)\|^2 \leqslant 2R^2\mathbb{E}\big[f_i(\theta) - f_i(\theta_*) - f_i'(\theta_*)^\top(\theta - \theta_*)\big]$ by the proof of co-coercivity, which is equal to $2R^2\big[g'(\theta) - g(\theta_*)\big]$

SVRG proof - from Bubeck (2015)

- **Lemma**: $\mathbb{E}\|f_i'(\theta) - f_i'(\theta_*)\|^2 \leqslant 2R^2\big[g(\theta) - g(\theta_*)\big]$

- From iteration $\theta_t = \theta_{t-1} - \gamma\big[f_{i_t}'(\theta_{t-1}) - f_{i_t}'(\tilde\theta) + g'(\tilde\theta)\big] = \theta_{t-1} - \gamma g_t$

$$\begin{aligned}
\|\theta_t - \theta_*\|^2 &= \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g_t + \gamma^2\|g_t\|^2 \\
\mathbb{E}\big[\|\theta_t - \theta_*\|^2|\mathcal{F}_{t-1}\big] &\leqslant \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\
&\quad + 2\gamma^2\|f_{i_t}'(\theta_{t-1}) - f_{i_t}'(\theta_*)\|^2 + 2\gamma^2\|f_{i_t}'(\tilde\theta) - f_{i_t}'(\theta_*) - g'(\tilde\theta)\|^2 \\
&\leqslant \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\
&\quad + 2\gamma^2R^2\big[g(\theta_{t-1}) - g(\theta_*) + g(\tilde\theta) - g(\theta_*)\big] \\
&\leqslant \|\theta_{t-1} - \theta_*\|^2 - 2\gamma(1 - 2\gamma R^2)[g(\theta_{t-1}) - g(\theta_*)] + 4R^2\gamma^2[g(\tilde\theta) - g(\theta_*)]
\end{aligned}$$

- By summing k times, we get:

$$\mathbb{E}\|\theta_k - \theta_*\|^2 \leqslant \|\theta_0 - \theta_*\|^2 - 2\gamma(1 - 2\gamma R^2)\sum_{t=1}^{k}\mathbb{E}[g(\theta_{t-1}) - g(\theta_*)] + 4kR^2\gamma^2[g(\tilde\theta) - g(\theta_*$$

which leads to the desired result

Fundamentals of constrained optimization

- We consider the following primal optimization problem

$$\min_{x \in D} f(x) \quad \text{s.t} \quad \forall i \in \{1, \ldots, m\}, h_i(x) = 0 \text{ and } \forall j \in \{1, \ldots, r\}, g_j(x) \leqslant 0$$

  - We denote by D^* the set of $x \in D$ satisfying the constraints

Fundamentals of constrained optimization

- We consider the following primal optimization problem

$$\min_{x \in D} f(x) \quad \text{s.t} \quad \forall i \in \{1, \dots, m\}, h_i(x) = 0 \text{ and } \forall j \in \{1, \dots, r\}, g_j(x) \leqslant 0$$

  – 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

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)$$

  – λ et μ 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): $\forall (\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^r_+$, $\forall x \in D^*$

$$\inf_{x' \in D} \mathcal{L}(x', \lambda, \mu) \leqslant \mathcal{L}(x, \lambda, \mu) \leqslant \sup_{(\lambda',\mu') \in \mathbb{R}^m \times \mathbb{R}^r_+} \mathcal{L}(x, \lambda', \mu')$$

which implies $q(\lambda, \mu) \leqslant f(x)$ and thus $d^* \leqslant p^*$

Sufficient conditions for strong duality

- **Geometric interpretation** for $\min_{x \in D} f(x)$ s.t $g(x) \leqslant 0$
  - Consider $A = \{(u,t) \in \mathbb{R}^2,\ \exists x \in D, f(x) \leqslant t, g(x) \leqslant u\}$

- **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-Kühn-Tucker (KKT) conditions**: If strong duality holds, then x^* is primal optimal and (λ^*, μ^*) are dual optimal if and only if:
  - *Primal stationarity*: x^* minimizes $x \mapsto \mathcal{L}(x, \lambda^*, \mu^*)$.
  - *Feasibility*: x^* and (λ^*, μ^*) 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**: $\min_{Ax=b, x \geqslant 0} c^\top x = \max_{A^\top y \leqslant c} b^\top y$

- **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**: $\min_{x^\top x \leqslant 1} \frac{1}{2} x^\top Q x - q^\top x$

Dual stochastic coordinate ascent - I

- **General learning formulation**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta^\top \Phi(x_i)) + \frac{\mu}{2} \|\theta\|_2^2$$

$$= \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \quad \frac{1}{n} \sum_{i=1}^{n} \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 \text{ such that } \forall i, u_i = \theta^\top \Phi(x_i)$$

$$= \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} \quad \frac{1}{n} \sum_{i=1}^{n} \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^{n} \alpha_i (u_i - \theta^\top \Phi(x_i))$$

$$= \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \quad \frac{1}{n} \sum_{i=1}^{n} \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^{n} \alpha_i (u_i - \theta^\top \Phi(x_i))$$

$$= \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \quad \frac{1}{n} \sum_{i=1}^{n} \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^{n} \alpha_i (u_i - \theta^\top \Phi(x_i))$$

Dual stochastic coordinate ascent - II

- **General learning formulation**:

$$
\begin{aligned}
& \min_{\theta \in \mathbb{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta^\top \Phi(x_i)) + \frac{\mu}{2} \|\theta\|_2^2 \\
= & \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathbb{R}^d, u \in \mathbb{R}^n} \ \frac{1}{n} \sum_{i=1}^{n} \ell_i(u_i) + \frac{\mu}{2} \|\theta\|_2^2 + \sum_{i=1}^{n} \alpha_i (u_i - \theta^\top \Phi(x_i)) \\
= & \max_{\alpha \in \mathbb{R}^n} \ \sum_{i=1}^{n} \max_{u_i \in \mathbb{R}} \left\{ \frac{1}{n} \ell_i(u_i) + \alpha_i u_i \right\} - \frac{1}{2\mu} \left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|_2^2 \\
= & \max_{\alpha \in \mathbb{R}^n} \ - \sum_{i=1}^{n} \psi_i(\alpha_i) - \frac{1}{2\mu} \left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|_2^2
\end{aligned}
$$

  – Minimizers obtained as $\theta = \frac{1}{\mu} \sum_{i=1}^{n} \alpha_i \Phi(x_i)$
  – ψ_i convex (up to affine transform = Fenchel-Legendre dual of ℓ_i)

Dual stochastic coordinate ascent - III

- **General learning formulation**:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta^\top \Phi(x_i)) + \frac{\mu}{2} \|\theta\|_2^2 = \max_{\alpha \in \mathbb{R}^n} \; - \sum_{i=1}^{n} \psi_i(\alpha_i) - \frac{1}{2\mu} \bigg\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \bigg\|_2^2$$

- **From primal to dual**
  - ℓ_i smooth $\Leftrightarrow$ ψ_i strongly convex
  - ℓ_i strongly convex $\Leftrightarrow$ ψ_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**: $\max_{\alpha \in \mathbb{R}^n} \ -\sum_{i=1}^{n} \psi_i(\alpha_i) - \frac{1}{2\mu}\Big\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \Big\|_2^2$

- **Stochastic coordinate descent**: at iteration t
  - Choose a coordinate i at random
  - Optimzte w.r.t. α_i: $\max_{\alpha_i \in \mathbb{R}} -\psi_i(\alpha_i) - \frac{1}{2\mu}\Big\| \alpha_i \Phi(x_i) + \sum_{j \neq i} \alpha_i \Phi(x_i) \Big\|_2^2$
  - 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 $f : \mathbb{R}^n \to \mathbb{R}$ which is L-smooth
  - Local smoothness constants $L_i = \sup_{\alpha \in \mathbb{R}^n} f''_{ii}(\alpha)$
  - $\max_{i \in \{1,\dots,n\}} L_i \leqslant L$ and $L \leqslant \sum_{i=1}^n L_i$
  - NB: in dual problems in machine learning $R^2 = \max_{i \in \{1,\dots,n\}} L_i$

- **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, $f(\alpha_t) \leqslant f(\alpha_{t-1}) - f'(\alpha_{t-1})^\top (\alpha_t - \alpha_{t-1}) + \frac{L_{i_t}}{2} \|\alpha_t - \alpha t - 1\|^2$ leading to $f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{1}{2L_{i_t}} |f'(\alpha_{t-1})_{i_t}|^2$
- Taking expectations: $\mathbb{E}[f(\alpha_t)|\mathcal{F}_{t-1}] \leqslant f(\alpha_{t-1}) - \sum_{i=1}^n \frac{p_i}{2L_i} |f'(\alpha_{t-1})_i|^2$

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, $f(\alpha_t) \leqslant f(\alpha_{t-1}) - f'(\alpha_{t-1})^\top(\alpha_t - \alpha_{t-1}) + \frac{L_{i_t}}{2}\|\alpha_t - \alpha t - 1\|^2$ leading to $f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{1}{2L_{i_t}}|f'(\alpha_{t-1})_{i_t}|^2$

- Taking expectations: $\mathbb{E}[f(\alpha_t)|\mathcal{F}_{t-1}] \leqslant f(\alpha_{t-1}) - \sum_{i=1}^n \frac{p_i}{2L_i}|f'(\alpha_{t-1})_i|^2$

- If $p_i = 1/n$ (uniform), $\mathbb{E}f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{1}{2n \max_i L_i}\|f'(\alpha_{t-1})\|^2$
  With strong convexity, this leads to $\mathbb{E}f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{\mu}{n \max_i L_i}\big[f(\alpha_{t-1} - f(\alpha^*)\big]$
  leading to a linear convergence rate with factor $1 - \dfrac{\mu}{n \max_i L_i}$

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, $f(\alpha_t) \leqslant f(\alpha_{t-1}) - f'(\alpha_{t-1})^\top (\alpha_t - \alpha_{t-1}) + \frac{L_{i_t}}{2} \|\alpha_t - \alpha t - 1\|^2$
  leading to $f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{1}{2L_{i_t}} |f'(\alpha_{t-1})_{i_t}|^2$

- Taking expectations: $\mathbb{E}[f(\alpha_t) | \mathcal{F}_{t-1}] \leqslant f(\alpha_{t-1}) - \sum_{i=1}^n \frac{p_i}{2L_i} |f'(\alpha_{t-1})_i|^2$

- If $p_i = 1/n$ (uniform), $\mathbb{E} f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{1}{2n \max_i L_i} \|f'(\alpha_{t-1})\|^2$
  With strong convexity, this leads to $\mathbb{E} f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{\mu}{n \max_i L_i} \big[f(\alpha_{t-1} - f(\alpha^*) \big]$
  leading to a linear convergence rate with factor $1 - \dfrac{\mu}{n \max_i L_i}$

- If $p_i = \frac{L_i}{\sum_{j=1}^n L_j}$, $\mathbb{E} f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{1}{2 \sum_{j=1}^n L_j} \|f'(\alpha_{t-1})\|^2$
  With strong convexity, this leads to $\mathbb{E} f(\alpha_t) \leqslant f(\alpha_{t-1}) - \frac{\mu}{\sum_{j=1}^n L_j} \big[f(\alpha_{t-1} - f(\alpha^*) \big]$
  leading to a linear convergence rate with factor $1 - \dfrac{\mu}{\sum_{j=1}^n L_j}$

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 $1 - \dfrac{\mu}{n \max_i L_i}$
  - If $p_i = \frac{L_i}{\sum_{j=1}^n L_j}$, linear rate $1 - \dfrac{\mu}{\sum_{j=1}^n L_j}$
- 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*

$$\begin{cases} \bar{\theta}_t \in \arg\min\limits_{\theta \in \mathcal{C}} \theta^\top f'(\theta_{t-1}) \\ \theta_t = (1 - \rho_t)\theta_{t-1} + \rho_t \bar{\theta}_t \end{cases}$$

Frank-Wolfe - conditional gradient - II

- **Convergence rates**: $f(\theta_t) - f(\theta_*) \leqslant \dfrac{2L\mathrm{diam}(\mathcal{C})^2}{t+1}$

Iteration: $\begin{cases} \bar\theta_t \in \arg\min_{\theta\in\mathcal{C}} \theta^\top f'(\theta_{t-1}) \\ \theta_t = (1-\rho_t)\theta_{t-1} + \rho_t\bar\theta_t \end{cases}$

From smoothness: $f(\theta_t) \leqslant f(\theta_{t-1}) + f'(\theta_{t-1})^\top\big[\rho_t(\bar\theta_t - \theta_{t-1})\big] + \dfrac{L}{2}\big\|\rho_t(\bar\theta_t - \theta_{t-1})\big\|^2$

From compactness: $f(\theta_t) \leqslant f(\theta_{t-1}) + f'(\theta_{t-1})^\top\big[\rho_t(\bar\theta_t - \theta_{t-1})\big] + \dfrac{L}{2}\rho_t^2\mathrm{diam}(\mathcal{C})^2$

From convexity, $f(\theta_t) - f(\theta_*) \leqslant f'(\theta_{t-1})^\top(\theta_{t-1} - \theta_*) \leqslant \max_{\theta\in\mathcal{C}} f'(\theta_{t-1})^\top(\theta_{t-1} - \theta)$,
which is equal to $f'(\theta_{t-1})^\top(\theta_{t-1} - \bar\theta_t -)$

Thus, $f(\theta_t) \leqslant f(\theta_{t-1}) - \rho_t\big[f(\theta_{t-1}) - f(\theta_*)\big] + \dfrac{L}{2}\rho_t^2\mathrm{diam}(\mathcal{C})^2$

With $\rho_t = 2/(t+1)$: $f(\theta_t) \leqslant \frac{2L\mathrm{diam}(\mathcal{C})^2}{t+1}$ *by direct expansion*

Frank-Wolfe - conditional gradient - II

- **Convergence rates**: $f(\theta_t) - f(\theta_*) \leqslant \dfrac{2L\text{diam}(\mathcal{C})^2}{t}$

- **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. **Classical stochastic approximation**

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

Outline - II

4. **Non-smooth stochastic approximation**
   - Stochastic (sub)gradient and averaging
   - Non-asymptotic results and lower bounds
   - Strongly convex vs. non-strongly convex

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 $\|\Phi(x)\|_2 \leqslant R$ 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 $\Theta = \{\|\theta\|_2 \leqslant D\}$

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

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

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

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

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

Stochastic subgradient “descent”/method

- **Assumptions**
  - f_n convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
  - (f_n) i.i.d. functions such that $\mathbb{E} f_n = f$
  - θ_* global optimum of f on $\{\|\theta\|_2 \leqslant D\}$
- **Algorithm**: $\theta_n = \Pi_D\Big(\theta_{n-1} - \frac{2D}{B\sqrt{n}} f_n'(\theta_{n-1})\Big)$
- **Bound**:

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

- “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: $O\Big(\dfrac{\sigma^2\gamma_n}{\mu}\Big) + O(e^{-\mu n\gamma_n})\|\theta_0 - \theta_*\|^2$
  - averaging: $\dfrac{\operatorname{tr} H(\theta_*)^{-1}}{n} + \mu^{-1}O(n^{-2\alpha} + n^{-2+\alpha}) + O\Big(\dfrac{\|\theta_0 - \theta_*\|^2}{\mu^2 n^2}\Big)$

Least-mean-square algorithm

- **Least-squares**: $f(\theta) = \frac{1}{2}\mathbb{E}\big[(y_n - \langle \Phi(x_n), \theta\rangle)^2\big]$ 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 $\mathbb{E}\big[\Phi(x_n) \otimes \Phi(x_n)\big] = H \succcurlyeq \mu \cdot \mathrm{Id}$

- **New analysis for averaging and constant step-size** $\gamma = 1/(4R^2)$
  - Assume $\|\Phi(x_n)\| \leqslant R$ and $|y_n - \langle \Phi(x_n), \theta_*\rangle| \leqslant \sigma$ almost surely
  - No assumption regarding lowest eigenvalues of H
  - Main result: $\boxed{\mathbb{E} f(\bar{\theta}_{n-1}) - f(\theta_*) \leqslant \frac{4\sigma^2 d}{n} + \frac{4R^2\|\theta_0 - \theta_*\|^2}{n}}$

- **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: $\boxed{\theta_n = \theta_{n-1} - \gamma\big[f_n'(\bar{\theta}_{n-1}) + f_n''(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1})\big]}$
  - 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
  - μ 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

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 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.

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.

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. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

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. ISSN 1935-8237. doi: 10.1561/2200000050. URL http://dx.doi.org/10.1561/2200000050.

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.

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

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, 2014.

A. Défossez and F. Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. 2015.

B. Delyon and A. Juditsky. Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3: 868–881, 1993.

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.

M. P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *arXiv:1104.2373*, 2011.

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.

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. Kesten. Accelerated stochastic approximation. *Ann. Math. Stat.*, 29(1):41–59, 1958.

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.

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.

O. Macchi. *Adaptive processing: The least mean squares approach with applications in transmission*. Wiley West Sussex, 1995.

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.

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. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851.

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 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, 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.

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. Sunehag, J. Trumpf, SVN Vishwanathan, and N. Schraudolph. Variable metric stochastic approximation theory. *International Conference on Artificial Intelligence and Statistics*, 2009.

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. 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.