Large-scale machine learning and convex optimization

Francis Bach
INRIA - Ecole Normale Supérieure, Paris, France

IFCAM, Bangalore - July 2014
“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

Google search for "fete de la science".
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

1. Large-scale machine learning and optimization
   - Traditional statistical analysis
   - Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   - Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
Supervised machine learning

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

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

convex data fitting term $+$ regularizer
Usual losses

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

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

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

![Graph showing various loss functions](image-url)
Main motivating examples

• **Support vector machine** (hinge loss)

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

• **Logistic regression**

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

• **Least-squares regression**

\[
\ell(Y, \theta^\top \Phi(X)) = \frac{1}{2}(Y - \theta^\top \Phi(X))^2
\]
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: \[ \theta = \sum_{i=1}^{n} \alpha_i \Phi(x_i) \]

- **Sparsity-inducing norms**
  - Main example: \( \ell_1 \)-norm \[ \|\theta\|_1 = \sum_{j=1}^{d} |\theta_j| \]
  - Perform model selection as well as regularization
  - Non-smooth optimization and structured sparsity
  - See, e.g., Bach, Jenatton, Mairal, and Obozinski (2011, 2012)
Supervised machine learning

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

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

convex data fitting term + regularizer
Supervised machine learning

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

- Prediction as a linear function $\theta^\top \Phi(x)$ of features $\Phi(x) \in \mathbb{R}^d$

- **(regularized) empirical risk minimization:** find $\hat{\theta}$ solution of

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

convex data fitting term $+$ regularizer

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$ training cost

- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$ testing cost

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

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

- Prediction as a linear function $\theta^\top \Phi(x)$ of features $\Phi(x) \in \mathbb{R}^d$

- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

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

convex data fitting term + regularizer

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^\top \Phi(x_i))$ training cost

- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$ testing cost

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

  - May be tackled simultaneously
Supervised machine learning

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

- Prediction as a linear function $\theta^T \Phi(x)$ of features $\Phi(x) \in \mathbb{R}^d$

- **(regularized) empirical risk minimization:** find $\hat{\theta}$ solution of

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

  convex data fitting term + constraint

- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^T \Phi(x_i))$ training cost

- Expected risk: $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^T \Phi(x))$ testing cost

- **Two fundamental questions:** (1) computing $\hat{\theta}$ and (2) analyzing $\hat{\theta}$
  
  - May be tackled simultaneously
General assumptions

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

- **Bounded features** $\Phi(x) \in \mathbb{R}^d$: $\|\Phi(x)\|_2 \leq R$

- **Empirical risk:** $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^\top \Phi(x_i))$ training cost

- **Expected risk:** $f(\theta) = \mathbb{E}_{(x,y)} \ell(y, \theta^\top \Phi(x))$ testing cost

- **Loss for a single observation:** $f_i(\theta) = \ell(y_i, \theta^\top \Phi(x_i))$
  \[ \Rightarrow \forall i, f(\theta) = \mathbb{E} f_i(\theta) \]

- **Properties of $f_i, f, \hat{f}$**
  - **Convex** on $\mathbb{R}^d$
  - Additional regularity assumptions: Lipschitz-continuity, smoothness and strong convexity
Lipschitz continuity

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

\[ \forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \implies \|f'(\theta)\|_2 \leq B \]

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

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

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

- If $f$ is twice differentiable: $\forall \theta \in \mathbb{R}^d, f''(\theta) \preceq L \cdot \text{Id}$
Smoothness and strong convexity

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

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

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

• Machine learning
  – with $f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^\top \Phi(x_i))$
  – Hessian $\approx$ covariance matrix $\frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^\top$
  – $\ell$-smooth loss and $R$-bounded data: $L = \ell R^2$
Smoothness and strong convexity

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

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

- If $f$ is twice differentiable: $\forall \theta \in \mathbb{R}^d, \ f''(\theta) \succeq \mu \cdot \text{Id}$
**Smoothness and strong convexity**

- A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is $\mu$-strongly convex if and only if
  \[
  \forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad f(\theta_1) \geq f(\theta_2) + f'(\theta_2)\top(\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2
  \]

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

- **Machine learning**
  - with $f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta\top \Phi(x_i))$
  - 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 \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is \( \mu \)-strongly convex if and only if

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

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

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

- Adding regularization by \( \frac{\mu}{2} \|\theta\|^2 \)
  - creates additional bias unless \( \mu \) is small
Summary of smoothness/convexity assumptions

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

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

- **Smoothness of $f$:** the function $f$ is convex, differentiable with $L$-Lipschitz-continuous gradient $f'$:

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

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

  \[ \forall \theta_1, \theta_2 \in \mathbb{R}^d, f(\theta_1) \geq f(\theta_2) + f'(\theta_2)^\top(\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2 \]
Analysis of empirical risk minimization

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

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

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

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

\[
f(\hat{\theta}) - \min_{\theta \in C} f(\theta) \leq 2 \sup_{\theta \in C} |\hat{f}(\theta) - f(\theta)| \quad (\text{proof})
\]

- Typically slow rate \( O\left(\frac{1}{\sqrt{n}}\right) \)

2. More refined concentration results with faster rates
Motivation from least-squares

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

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

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

\[
\sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| \leq O\left(\frac{1}{\sqrt{n}}\right) \text{ with high probability}
\]
Slow rate for supervised learning

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

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

- With probability greater than $1 - \delta$
  
  $$
  \sup_{\theta \in C} |\hat{f}(\theta) - f(\theta)| \leq \frac{GRD}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]
  $$

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

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

- **Lipschitz functions $\Rightarrow$ slow rate**
Symmetrization with Rademacher variables

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

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

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

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

- Note two expectations, with respect to \(D\) and with respect to \(\varepsilon\)

- Main property:

\[
\mathbb{E} \left[ \sup_{\theta \in \Theta} \left| f(\theta) - \hat{f}(\theta) \right| \right] \leq 2R_n
\]
From Rademacher complexity to uniform bound

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

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

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

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

  \[
  Z \leq E_Z + \sqrt{\frac{n}{2} c \cdot \sqrt{\log \frac{1}{\delta}}} \leq 2R_n + \frac{\sqrt{2}}{\sqrt{n}} (\ell_0 + GRD) \sqrt{\log \frac{1}{\delta}}
  \]
Bounding the Rademacher average - I

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

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

$$\leq \mathbb{E} \left[ \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_i(0) \right| \right] + \mathbb{E} \left[ \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left[ f_i(\theta) - f_i(0) \right] \right| \right]$$

$$\leq \frac{\ell_0}{\sqrt{n}} + \mathbb{E} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left[ f_i(\theta) - f_i(0) \right] \right]$$

$$= \frac{\ell_0}{\sqrt{n}} + \mathbb{E} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \varphi_i(\theta^\top \Phi(x_i)) \right]$$

- Using Ledoux-Talagrand concentration results for Rademacher averages (since $\varphi_i$ is $G$-Lipschitz, we get:

$$R_n \leq \frac{\ell_0}{\sqrt{n}} + 2G \cdot \mathbb{E} \left[ \sup_{\|\theta\|_2 \leq D} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \theta^\top \Phi(x_i) \right| \right]$$
Bounding the Rademacher average - II

- We have:

\[
R_n \leq \frac{\ell_0}{\sqrt{n}} + 2G \mathbb{E} \left[ \sup_{\|\theta\|_2 \leq D} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \theta^\top \Phi(x_i) \right\| \right]
\]

\[
= \frac{\ell_0}{\sqrt{n}} + 2G \mathbb{E} \left[ D \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Phi(x_i) \right]_2
\]

\[
\leq \frac{\ell_0}{\sqrt{n}} + 2GD \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Phi(x_i) \right]_2^2
\]

\[
\leq \frac{2(\ell_0 + GRD)}{\sqrt{n}}
\]

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

\[
\sup_{\theta \in \Theta} |f(\theta) - \hat{f}(\theta)| \leq \frac{1}{\sqrt{n}} (\ell_0 + GRD) (4 + \sqrt{2 \log \frac{1}{\delta}})
\]
Putting it all together

- We have, with probability $1 - \delta$, for all $\theta \in \Theta$:

\[
 f(\theta) - f(\theta^*) \leq \left[ f(\theta) - \hat{f}(\theta) \right] + \left[ \hat{f}(\theta) - \min_{\theta' \in \Theta} \hat{f}(\theta') \right] + \left[ \min_{\theta' \in \Theta} \hat{f}(\theta') - \hat{f}(\theta^*) \right]
\]

\[
 \leq \frac{2}{\sqrt{n}} (\ell_0 + GRD) (4 + \sqrt{2 \log \frac{1}{\delta}}) + \left[ \hat{f}(\theta) - \min_{\theta' \in \Theta} \hat{f}(\theta') \right]
\]

- Only need to optimize with precision $\frac{2}{\sqrt{n}} (\ell_0 + GRD)$
**Slow rate for supervised learning (summary)**

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

- With probability greater than $1 - \delta$
  \[
  \sup_{\theta \in C} |\hat{f}(\theta) - f(\theta)| \leq \frac{(\ell_0 + GRD)}{\sqrt{n}} \left[ 2 + \sqrt{2 \log \frac{2}{\delta}} \right]
  \]

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

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

- **Lipschitz functions $\Rightarrow$ slow rate**
Motivation from mean estimation

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

- From before:
  
  - \( f(\theta) = \frac{1}{2} \mathbb{E} (\theta - z)^2 = \frac{1}{2} (\theta - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z) = \hat{f}(\theta) + O(1/\sqrt{n}) \)
  
  - \( f(\hat{\theta}) = \frac{1}{2} (\hat{\theta} - \mathbb{E}z)^2 + \frac{1}{2} \text{var}(z) = f(\mathbb{E}z) + O(1/\sqrt{n}) \)
Motivation from mean estimation

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

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

- More refined/direct bound:
  \[ f(\hat{\theta}) - f(\mathbb{E}z) = \frac{1}{2} (\hat{\theta} - \mathbb{E}z)^2 \]
  \[ \mathbb{E} [f(\hat{\theta}) - f(\mathbb{E}z)] = \frac{1}{2} \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} z_i - \mathbb{E}z \right)^2 = \frac{1}{2n} \text{var}(z) \]

- Bound only at $\hat{\theta}$ + strong convexity
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: \(f^\mu(\theta) = f(\theta) + \frac{\mu}{2}\|\theta\|^2\) and \(\hat{f}^\mu(\theta) = \hat{f}(\theta) + \frac{\mu}{2}\|\theta\|^2\)
  - Convexity

- For any \(a > 0\), with probability greater than \(1 - \delta\), for all \(\theta \in \mathbb{R}^d\),
  \[
  f^\mu(\theta) - \min_{\eta \in \mathbb{R}^d} f^\mu(\eta) \leq (1 + a)(\hat{f}^\mu(\theta) - \min_{\eta \in \mathbb{R}^d} \hat{f}^\mu(\eta)) + \frac{8(1 + \frac{1}{a})G^2 R^2 (32 + \log \frac{1}{\delta})}{\mu n}
  \]

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

- **Strongly convex functions \(\Rightarrow\) fast rate
  
  - Warning: \(\mu\) should decrease with \(n\) to reduce approximation error
Outline

1. Large-scale machine learning and optimization
   - Traditional statistical analysis
   - Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   - Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
Complexity results in convex optimization

- **Assumption**: $f$ convex on $\mathbb{R}^d$

- **Classical generic algorithms**
  - (sub)gradient method/descent
  - Accelerated gradient descent
  - Newton method

- **Key additional properties of $f$**
  - Lipschitz continuity, smoothness or strong convexity

- **Key insight from Bottou and Bousquet (2008)**
  - In machine learning, no need to optimize below estimation error

- **Key reference**: Nesterov (2004)
Subgradient method/descent

• Assumptions
  – $f$ convex and $B$-Lipschitz-continuous on $\{\|\theta\|_2 \leq D\}$

• Algorithm:
  $\theta_t = \Pi_D \left( \theta_{t-1} - \frac{2D}{B\sqrt{t}} f'(\theta_{t-1}) \right)$
  – $\Pi_D$ : orthogonal projection onto $\{\|\theta\|_2 \leq D\}$

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

• Three-line proof

• Best possible convergence rate after $O(d)$ iterations
Subgradient method/descent - proof - 1

• Iteration: $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t f'(\theta_{t-1}))$ with $\gamma_t = \frac{2D}{B\sqrt{t}}$

• Assumption: $\|f'(\theta)\|_2 \leq B$ and $\|\theta\|_2 \leq D$

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

• leading to

\[ f(\theta_{t-1}) - f(\theta_*) \leq \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}[[\|\theta_{t-1} - \theta_*\|^2_2 - \|\theta_t - \theta_*\|^2_2]] \]
Subgradient method/descent - proof - II

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

\[
\sum_{u=1}^{t} \left[ f(\theta_{u-1}) - f(\theta_*) \right] \leq \sum_{u=1}^{t} \frac{B^2\gamma_u}{2} + \sum_{u=1}^{t} \frac{1}{2\gamma_u} \left[ \|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2 \right]
\]

\[
= \sum_{u=1}^{t} \frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1} \|\theta_u - \theta_*\|_2^2 \left( \frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u} \right) + \frac{\|\theta_0 - \theta_*\|_2^2}{2\gamma_1} - \frac{\|\theta_t - \theta_*\|_2^2}{2\gamma_t}
\]

\[
\leq \sum_{u=1}^{t} \frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1} 4D^2 \left( \frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u} \right) + \frac{4D^2}{2\gamma_1}
\]

\[
= \sum_{u=1}^{t} \frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_t} \leq 2DB\sqrt{t} \quad \text{with} \quad \gamma_t = \frac{2D}{B\sqrt{t}}
\]

- Using convexity: \( f\left( \frac{1}{t} \sum_{k=0}^{t-1} \theta_k \right) - f(\theta_*) \leq \frac{2DB}{\sqrt{t}} \)
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 \(\|\Phi(x)\|_2 \leq R\) a.s.
  - \(\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \Phi(x_i)^\top \theta)\)
  - G-Lipschitz loss: \(f\) and \(\hat{f}\) are \(GR\)-Lipschitz on \(C = \{\|\theta\|_2 \leq D\}\)

- **Statistics:** with probability greater than \(1 - \delta\)
  \[
  \sup_{\theta \in C} |\hat{f}(\theta) - f(\theta)| \leq \frac{GRD}{\sqrt{n}} \left[2 + \sqrt{2 \log \frac{2}{\delta}}\right]
  \]

- **Optimization:** after \(t\) iterations of subgradient method
  \[
  \hat{f}(\hat{\theta}) - \min_{\eta \in C} \hat{f}(\eta) \leq \frac{GRD}{\sqrt{t}}
  \]

- \(t = n\) iterations, with total running-time complexity of \(O(n^2d)\)
Subgradient descent - strong convexity

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

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

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

- **Three-line proof**

- **Best possible convergence rate after $O(d)$ iterations**
Subgradient method - strong convexity - proof - I

• Iteration: \( \theta_t = \Pi_D(\theta_{t-1} - \gamma_t f'(\theta_{t-1})) \) with \( \gamma_t = \frac{2}{\mu(t+1)} \)

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

\[
\|\theta_t - \theta_*\|_2^2 \leq \|\theta_{t-1} - \theta_* - \gamma_t f'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections}
\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2 \gamma_t^2 - 2 \gamma_t (\theta_{t-1} - \theta_*)^\top f'(\theta_{t-1}) \text{ because } \|f'(\theta_{t-1})\|_2 \leq B
\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2 \gamma_t^2 - 2 \gamma_t \left[ f(\theta_{t-1}) - f(\theta_*) + \frac{\mu}{2} \|\theta_{t-1} - \theta_*\|_2^2 \right]
\]

(property of subgradients and strong convexity)

• leading to

\[
f(\theta_{t-1}) - f(\theta_*) \leq \frac{B^2 \gamma_t}{2} + \frac{1}{2} \left[ \frac{1}{\gamma_t} - \mu \right] \|\theta_{t-1} - \theta_*\|_2^2 - \frac{1}{2 \gamma_t} \|\theta_t - \theta_*\|_2^2 \leq \frac{B^2}{\mu(t+1)} + \frac{\mu}{2} \left[ \frac{t-1}{2} \right] \|\theta_{t-1} - \theta_*\|_2^2 - \frac{\mu(t+1)}{4} \|\theta_t - \theta_*\|_2^2
\]
Subgradient method - strong convexity - proof - II

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

\[
\sum_{u=1}^{t} u [f(\theta_{u-1}) - f(\theta_*)] \leq \sum_{t=1}^{u} \frac{B^2u}{\mu(u + 1)} + \frac{1}{4} \sum_{u=1}^{t} [u(u - 1)\|\theta_{u-1} - \theta_*\|^2 - u(u + 1)\|\theta_u - \theta_*\|^2] \\
\leq \frac{B^2t}{\mu} + \frac{1}{4} [0 - t(t + 1)\|\theta_t - \theta_*\|^2] \leq \frac{B^2t}{\mu}
\]

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

- **Assumptions**
  - $f$ convex with $L$-Lipschitz-continuous gradient
  - Minimum attained at $\theta^*$

- **Algorithm:**
  \[
  \theta_t = \theta_{t-1} - \frac{1}{L} f'(\theta_{t-1})
  \]

- **Bound:**
  \[
  f(\theta_t) - f(\theta^*) \leq \frac{2L \|\theta_0 - \theta^*\|^2}{t + 4}
  \]

- **Three-line proof**

- **Not best possible convergence rate after $O(d)$ iterations**
(smooth) gradient descent - strong convexity

- **Assumptions**
  - $f$ convex with $L$-Lipschitz-continuous gradient
  - $f$ $\mu$-strongly convex

- **Algorithm**:
  $$\theta_t = \theta_{t-1} - \frac{1}{L} f' (\theta_{t-1})$$

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

- **Three-line proof**

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

- **Line search**
Accelerated gradient methods (Nesterov, 1983)

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

• **Algorithm:**
  
  \[
  \theta_t = \eta_{t-1} - \frac{1}{L}f'((\eta_{t-1})
  \]
  
  \[
  \eta_t = \theta_t + \frac{t-1}{t+2}(\theta_t - \theta_{t-1})
  \]

• **Bound:**
  
  \[
  f(\theta_t) - f(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{(t + 1)^2}
  \]

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

• Not improvable

• Extension to strongly convex functions
Optimization for sparsity-inducing norms (see Bach, Jenatton, Mairal, and Obozinski, 2011)

- 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, 2011)

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

- Problems of the form:
  \[
  \min_{\theta \in \mathbb{R}^d} f(\theta) + \mu \Omega(\theta)
  \]

  \[
  \begin{align*}
  \theta_{t+1} &= \arg \min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \mu \Omega(\theta) + \frac{L}{2} \|\theta - \theta_t\|^2_2 \\
  \Omega(\theta) &= \|\theta\|_1 \Rightarrow \text{Thresholded gradient descent}
  \end{align*}
  \]

- Similar convergence rates than smooth optimization
  \[
  \begin{align*}
  \text{Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)}
  \end{align*}
  \]
Summary: minimizing convex functions

- **Assumption:** \( f \) convex

- **Gradient descent:** \( \theta_t = \theta_{t-1} - \gamma_t f'(\theta_{t-1}) \)
  - \( O(1/\sqrt{t}) \) 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:** \( \theta_t = \theta_{t-1} - f'''(\theta_{t-1})^{-1} f'(\theta_{t-1}) \)
  - \( O(e^{-\rho^2 t}) \) convergence rate
Summary: minimizing convex functions

• Assumption: \( f \) convex

• Gradient descent: \( \theta_t = \theta_{t-1} - \gamma_t f'(\theta_{t-1}) \)
  
  – \( O(1/\sqrt{t}) \) 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: \( \theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1}) \)
  
  – \( O(e^{-\rho^2 t}) \) convergence rate

• Key insights from Bottou and Bousquet (2008)
  1. In machine learning, no need to optimize below statistical error
  2. In machine learning, cost functions are averages

⇒ Stochastic approximation
Outline

1. Large-scale machine learning and optimization
   - Traditional statistical analysis
   - Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   - Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
Stochastic approximation

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

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

• **Machine learning - statistics**

  – loss for a single pair of observations: $f_n(\theta) = \ell(y_n, \theta^\top \Phi(x_n))$

  – $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E} \ell(y_n, \theta^\top \Phi(x_n)) = \text{generalization error}$

  – Expected gradient: $f'(\theta) = \mathbb{E} f'_n(\theta) = \mathbb{E} \left\{ \ell'(y_n, \theta^\top \Phi(x_n)) \Phi(x_n) \right\}$

  – Non-asymptotic results

• **Number of iterations = number of observations**
Stochastic approximation

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

- **Stochastic approximation**
  - (much) broader applicability beyond convex optimization

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

  - Beyond convex problems, i.i.d assumption, finite dimension, etc.
  - Typically asymptotic results
  - See, e.g., Kushner and Yin (2003); Borkar (2008); Benveniste et al. (2012)
Relationship to online learning

- **Stochastic approximation**
  - Minimize \( f(\theta) = \mathbb{E}_z \ell(\theta, z) = \text{generalization error} \) of \( \theta \)
  - Using the gradients of single i.i.d. observations
Relationship to online learning

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

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

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

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

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

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

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

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

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

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

– Which learning rate sequence $\gamma_n$? Classical setting: $\gamma_n = C n^{-\alpha}$
Convex stochastic approximation

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

- **Key algorithm:** Stochastic gradient descent (a.k.a. Robbins-Monro)
  \[
  \theta_n = \theta_{n-1} - \gamma_n f'_n(\theta_{n-1})
  \]
  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k$
  - Which learning rate sequence $\gamma_n$? Classical setting: $\gamma_n = Cn^{-\alpha}$

- **Desirable practical behavior**
  - Applicable (at least) to classical supervised learning problems
  - Robustness to (potentially unknown) constants ($L,B,\mu$)
  - Adaptivity to difficulty of the problem (e.g., strong convexity)
**Stochastic subgradient descent/method**

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

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

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

- “Same” three-line proof as in the deterministic case

- **Minimax convergence rate**

- **Running-time complexity:** $O(dn)$ after $n$ iterations
Stochastic subgradient method - proof - 1

• Iteration: \( \theta_n = \Pi_D(\theta_{n-1} - \gamma_n f'_n(\theta_{n-1})) \) with \( \gamma_n = \frac{2D}{B \sqrt{n}} \)

• \( F_n \): information up to time \( n \)

• \( \|f'_n(\theta)\|_2 \leq B \) and \( \|\theta\|_2 \leq D \), unbiased gradients/functions \( \mathbb{E}(f_n|F_{n-1}) = f \)

\[
\|\theta_n - \theta^*\|_2^2 \leq \|\theta_{n-1} - \theta^* - \gamma_n f'_n(\theta_{n-1})\|_2^2 \text{ by contractivity of projections}
\leq \|\theta_{n-1} - \theta^*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n (\theta_{n-1} - \theta^*)^\top f'_n(\theta_{n-1}) \text{ because } \|f'_n(\theta_{n-1})\|_2 \leq B
\]

\[
\mathbb{E}\left[\|\theta_n - \theta^*\|_2^2|F_{n-1}\right] \leq \|\theta_{n-1} - \theta^*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n (\theta_{n-1} - \theta^*)^\top f'(\theta_{n-1})
\leq \|\theta_{n-1} - \theta^*\|_2^2 + B^2 \gamma_n^2 - 2\gamma_n \left[ f(\theta_{n-1}) - f(\theta^*) \right] \text{ (subgradient property)}
\]

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

• leading to \( \mathbb{E}f(\theta_{n-1}) - f(\theta^*) \leq \frac{B^2 \gamma_n}{2} + \frac{1}{2\gamma_n} \left[ \mathbb{E}\|\theta_{n-1} - \theta^*\|_2^2 - \mathbb{E}\|\theta_n - \theta^*\|_2^2 \right] \)
Stochastic subgradient method - proof - II

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

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

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

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

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

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

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

- “Same” three-line proof than in the deterministic case

- **Minimax convergence rate**
Stochastic subgradient descent - strong convexity - II

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

• Algorithm:

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

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

• Minimax convergence rate
Outline

1. Large-scale machine learning and optimization
   - Traditional statistical analysis
   - Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   - Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
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 $\gamma_n = C n^{-\alpha}$ with $\alpha \in (1/2, 1)$ lead to $O(n^{-1})$ for smooth strongly convex problems
Convex stochastic approximation

Existing work

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

• Asymptotic analysis of averaging (Polyak and Juditsky, 1992; Ruppert, 1988)
  – All step sizes $\gamma_n = Cn^{-\alpha}$ with $\alpha \in (1/2, 1)$ lead to $O(n^{-1})$ for smooth strongly convex problems

• Non-asymptotic analysis for smooth problems?
Smoothness/convexity assumptions

• Iteration: \( \theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1}) \)

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

• Smoothness of \( f_n \): For each \( n \geq 1 \), the function \( f_n \) is a.s. convex, differentiable with \( L \)-Lipschitz-continuous gradient \( f_n' \):

  – Smooth loss and bounded data

• Strong convexity of \( f \): The function \( f \) is strongly convex with respect to the norm \( \| \cdot \| \), with convexity constant \( \mu > 0 \):

  – Invertible population covariance matrix
  – or regularization by \( \frac{\mu}{2} \| \theta \|^2 \)
Summary of new results (Bach and Moulines, 2011)

• Stochastic gradient descent with learning rate $\gamma_n = C n^{-\alpha}$

• **Strongly convex smooth objective functions**
  - Old: $O(n^{-1})$ rate achieved **without** averaging for $\alpha = 1$
  - New: $O(n^{-1})$ rate achieved **with** averaging for $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants
  - Forgetting of initial conditions
  - Robustness to the choice of $C$
Summary of new results (Bach and Moulines, 2011)

• Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$

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

• Convergence rates for $\mathbb{E}\|\theta_n - \theta^*\|^2$ and $\mathbb{E}\|\bar{\theta}_n - \theta^*\|^2$
  – no averaging: $O\left(\frac{\sigma^2 \gamma_n}{\mu}\right) + O(e^{-\mu n \gamma_n})\|\theta_0 - \theta^*\|^2$
  – averaging: $\frac{\text{tr} \, H(\theta^*)^{-1}}{n} + \mu^{-1}O(n^{-2\alpha} + n^{-2+\alpha}) + O\left(\frac{\|\theta_0 - \theta^*\|^2}{\mu^2 n^2}\right)$
Classical proof sketch (no averaging)

\[ \|\theta_n - \theta_*\|_2^2 = \|\theta_{n-1} - \gamma_n f'_n(\theta_{n-1}) - \theta_*\|_2^2 \]

\[ = \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) + \gamma_n^2 \|f'_n(\theta_{n-1})\|_2^2 \]

\[ \leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \]

\[ + 2\gamma_n^2 \|f'_n(\theta_*)\|_2^2 + 2\gamma_n^2 \|f'_n(\theta_{n-1}) - f'_n(\theta_*)\|^2 \]

\[ \leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \]

\[ + 2\gamma_n^2 \|f'_n(\theta_*)\|_2^2 + 2\gamma_n^2 L [f'_n(\theta_{n-1}) - f'_n(\theta_*)]^\top (\theta_{n-1} - \theta_*) \]

\[ \leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n (\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) \]

\[ + 2\gamma_n^2 \|f'_n(\theta_*)\|_2^2 + 2\gamma_n^2 L [f'(\theta_{n-1}) - 0]^\top (\theta_{n-1} - \theta_*) \]

\[ \leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n(1 - \gamma_n L)(\theta_{n-1} - \theta_*)^\top f'_n(\theta_{n-1}) + 2\gamma_n^2 \sigma^2 \]

\[ \leq \|\theta_{n-1} - \theta_*\|_2^2 - 2\gamma_n(1 - \gamma_n L) \frac{1}{2} \mu \|\theta_{n-1} - \theta_*\|_2^2 + 2\gamma_n^2 \sigma^2 \]

\[ = [1 - \mu \gamma_n (1 - \gamma_n L)] \|\theta_{n-1} - \theta_*\|_2^2 + 2\gamma_n^2 \sigma^2 \]

\[ \mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] \leq [1 - \mu \gamma_n (1 - \gamma_n L)] \mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] + 2\gamma_n^2 \sigma^2 \]
Proof sketch (averaging)

• From Polyak and Juditsky (1992):

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

\[\iff f'_n(\theta_{n-1}) = \frac{1}{\gamma_n} (\theta_{n-1} - \theta_n)\]

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

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

\[+ O(\|\theta_{n-1} - \theta_*\|) \varepsilon_n\]

\[\iff \theta_{n-1} - \theta_* = -f''_n(\theta_*)^{-1} f'_n(\theta_*) + \frac{1}{\gamma_n} f''_n(\theta_*)^{-1} (\theta_{n-1} - \theta_n)\]

\[+ O(\|\theta_{n-1} - \theta_*\|^2) + O(\|\theta_{n-1} - \theta_*\|) \varepsilon_n\]

• Averaging to cancel the term \(\frac{1}{\gamma_n} f''_n(\theta_*)^{-1} (\theta_{n-1} - \theta_n)\)
Robustness to wrong constants for $\gamma_n = C n^{-\alpha}$

- $f(\theta) = \frac{1}{2}|\theta|^2$ 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
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)
Outline

1. Large-scale machine learning and optimization
   • Traditional statistical analysis
   • Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   • Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
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 ⇒ local strong convexity**
  - unless restricted to \(|\theta^\top \Phi(x_n)| \leq M\) (and with constants \(e^M\))
  - \(\mu = \) lowest eigenvalue of the Hessian at the optimum \(f''(\theta^*)\)
Adaptive algorithm for logistic regression

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

- **Cannot be strongly convex** \(\Rightarrow\) **local** strong convexity
  - unless restricted to \(|\theta^\top \Phi(x_n)| \leq M\) (and with constants \(e^M\))
  - \(\mu = \) lowest eigenvalue of the Hessian at the optimum \(f''(\theta_*)\)

- **\(n\) steps of averaged SGD with constant step-size** \(1/(2R^2 \sqrt{n})\)
  - with \(R = \) radius of data (Bach, 2013):
    \[
    \mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \min \left\{ \frac{1}{\sqrt{n}}, \frac{R^2}{n\mu} \right\} (15 + 5R\|\theta_0 - \theta_*\|)^4
    \]
  - Proof based on self-concordance (Nesterov and Nemirovski, 1994)
Self-concordance

• Usual definition for convex $\varphi : \mathbb{R} \to \mathbb{R}$: $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$
  
  – Affine invariant
  
  – Extendable to all convex functions on $\mathbb{R}^d$ by looking at rays
  
  – Used for the sharp proof of quadratic convergence of Newton method (Nesterov and Nemirovski, 1994)

• Generalized notion: $|\varphi'''(t)| \leq \varphi''(t)$
  
  – Applicable to logistic regression (with extensions)
Self-concordance

- Usual definition for convex $\varphi : \mathbb{R} \to \mathbb{R}$: $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$
  - Affine invariant
  - Extendable to all convex functions on $\mathbb{R}^d$ by looking at rays
  - Used for the sharp proof of quadratic convergence of Newton method (Nesterov and Nemirovski, 1994)

- Generalized notion: $|\varphi'''(t)| \leq \varphi''(t)$
  - Applicable to logistic regression (with extensions)

- Important properties
  - Allows global Taylor expansions
  - Relates expansions of derivatives of different orders
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 2: \[ f'(\theta_{n-1}) = \frac{1}{\gamma}(\theta_{n-1} - \theta_n) \implies \frac{1}{n} \sum_{k=1}^{n} f'_k(\theta_{k-1}) = \frac{1}{n\gamma}(\theta_0 - \theta_n) \]

• Step 3: \[
  \left\| f'(\frac{1}{n} \sum_{k=1}^{n} \theta_{k-1}) - \frac{1}{n} \sum_{k=1}^{n} f'(\theta_{k-1}) \right\|_2 \\
  = O(f(\bar{\theta}_n) - f(\theta_*)) = O(1/\sqrt{n}) \text{ using self-concordance}
\]

• Step 4a: if \( f \) \( \mu \)-strongly convex, \[
  f(\bar{\theta}_n) - f(\theta_*) \leq \frac{1}{2\mu} \|f'(\bar{\theta}_n)\|_2^2
\]

• Step 4b: if \( f \) self-concordant, “locally true” with \( \mu = \lambda_{\text{min}}(f''(\theta_*)) \)
Adaptive algorithm for logistic regression

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

- **Cannot be strongly convex \(\Rightarrow\) local strong convexity**
  
  - unless restricted to \(|\theta^\top \Phi(x_n)| \leq M\) (and with constants \(e^M\))
  
  - \(\mu = \) lowest eigenvalue of the Hessian at the optimum \(f''(\theta_*)\)

- **\(n\) steps of averaged SGD with constant step-size** \(1/(2R^2 \sqrt{n})\)
  
  - with \(R = \) radius of data (Bach, 2013):
    
    \[
    \mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \min \left\{ \frac{1}{\sqrt{n}}, \frac{R^2}{n\mu} \right\} (15 + 5R\|\theta_0 - \theta_*\|)^4
    \]

  - Proof based on self-concordance (Nesterov and Nemirovski, 1994)
Adaptive algorithm for logistic regression

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

• Cannot be strongly convex \(\Rightarrow\) local strong convexity
  – unless restricted to \(|\theta^\top \Phi(x_n)| \leq M\) (and with constants \(e^M\))
  – \(\mu\) = lowest eigenvalue of the Hessian at the optimum \(f''''(\theta_*)\)

• \(n\) steps of averaged SGD with constant step-size \(1/(2R^2 \sqrt{n})\)
  – with \(R = \) radius of data (Bach, 2013):

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

– A single adaptive algorithm for smooth problems with convergence rate \(O(1/n)\) in all situations?
Least-mean-square algorithm

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

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

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

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

- LMS recursion:
  \[ \theta_n - \theta_* = \left( I - \gamma \Phi(x_n) \otimes \Phi(x_n) \right) (\theta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n) \]

- Simplified LMS recursion: with \( H = \mathbb{E}\left[ \Phi(x_n) \otimes \Phi(x_n) \right] \)
  \[ \theta_n - \theta_* = \left( I - \gamma H \right) (\theta_{n-1} - \theta_*) + \gamma \varepsilon_n \Phi(x_n) \]
  - Direct proof technique of Polyak and Juditsky (1992), e.g.,
  \[ \theta_n - \theta_* = \left( I - \gamma H \right)^n (\theta_0 - \theta_*) + \gamma \sum_{k=1}^{n} \left( I - \gamma H \right)^{n-k} \varepsilon_k \Phi(x_k) \]

- Infinite expansion of Aguech, Moulines, and Priouret (2000) in powers of \( \gamma \)
Markov chain interpretation of constant step sizes

- LMS recursion for $f_n(\theta) = \frac{1}{2}(y_n - \langle \Phi(x_n), \theta \rangle)^2$
  \[ \theta_n = \theta_{n-1} - \gamma(\langle \Phi(x_n), \theta_{n-1} \rangle - y_n)\Phi(x_n) \]

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

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

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

- For least-squares, $\bar{\theta}_\gamma = \theta^*$
Markov chain interpretation of constant step sizes

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

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

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

- For least-squares, $\bar{\theta}_\gamma = \theta^*$
Markov chain interpretation of constant step sizes

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

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

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

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

- Gaussian distributions - $p = 20$

![Graph showing synthetic square distribution](image)
Simulations - benchmarks

- *alpha* \((p = 500, \, n = 500 \, 000)\), *news* \((p = 1 \, 300 \, 000, \, n = 20 \, 000)\)
Beyond least-squares - Markov chain interpretation

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

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

- $\theta_n$ oscillates around the wrong value $\overline{\theta}_\gamma \neq \theta_*$
Beyond least-squares - Markov chain interpretation

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

\( \theta_n \) oscillates around the wrong value \( \bar{\theta}_\gamma \neq \theta_* \)
- moreover, \( \|\theta_* - \theta_n\| = O_p(\sqrt{\gamma}) \)

- 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 - $p = 20$

![Graph showing synthetic logistic - 1 with various $R^2$ values and log scale for $n$ and $f(\theta) - f(\theta_*)$]
Restoring convergence through online Newton steps

- **Known facts**

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

• Known facts

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

• Online Newton step

  – Rate: $O(((n^{-1/2})^2 + n^{-1}) = O(n^{-1})$
  – Complexity: $O(p)$ per iteration
The Newton step for \( f = \mathbb{E}f_n(\theta) \overset{\text{def}}{=} \mathbb{E}[\ell(y_n, \langle \theta, \Phi(x_n) \rangle)] \) at \( \tilde{\theta} \) is equivalent to minimizing the quadratic approximation

\[
g(\theta) = f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta}) \rangle
\]

\[
= f(\tilde{\theta}) + \langle \mathbb{E}f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, \mathbb{E}f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle
\]

\[
= \mathbb{E} \left[ f(\tilde{\theta}) + \langle f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2} \langle \theta - \tilde{\theta}, f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle \right]
\]
Restoring convergence through online Newton steps

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

$$g(\theta) = f(\tilde{\theta}) + \langle f'(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2}\langle \theta - \tilde{\theta}, f''(\tilde{\theta})(\theta - \tilde{\theta}) \rangle$$

$$= f(\tilde{\theta}) + \langle \mathbb{E}f'_n(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{1}{2}\langle \theta - \tilde{\theta}, \mathbb{E}f''_n(\tilde{\theta})(\theta - \tilde{\theta}) \rangle$$

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

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

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

- $f''_n(\tilde{\theta}) = \ell''(y_n, \langle \tilde{\theta}, \Phi(x_n) \rangle)\Phi(x_n) \otimes \Phi(x_n)$ has rank one

- New online Newton step without computing/inverting Hessians
Choice of support point for online Newton step

- **Two-stage procedure**
  
  (1) Run $n/2$ iterations of averaged SGD to obtain $\tilde{\theta}$
  
  (2) Run $n/2$ iterations of averaged constant step-size LMS

  - Reminiscent of one-step estimators (see, e.g., Van der Vaart, 2000)
  - Provable convergence rate of $O(p/n)$ for logistic regression
  - Additional assumptions but no strong convexity
Logistic regression - Proof technique

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

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

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

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

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

$$\mathbb{E}\langle \Phi(x_n), \eta \rangle^4 \leq \kappa \left[ \mathbb{E}\langle \Phi(x_n), \eta \rangle^2 \right]^2$$
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 $O(p/n)$** for logistic regression
     - Additional assumptions but no strong convexity

- **Update at each iteration using the current averaged iterate**
  - Recursion:  
    $$\theta_n = \theta_{n-1} - \gamma \left[ f_n'(\bar{\theta}_{n-1}) + f_n''(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1}) \right]$$
  - 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

\[
\begin{align*}
\theta_n & = \theta_{n-1} - \gamma \left[ f'_n(\bar{\theta}_{n-1}) + f''_n(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1}) \right] \\
\bar{\theta}_n & = \bar{\theta}_{n-1} + \frac{1}{n}(\theta_n - \bar{\theta}_{n-1})
\end{align*}
\]

• Instance of two-time-scale stochastic approximation (Borkar, 1997)
  
  – Given $\bar{\theta}$, $\theta_n = \theta_{n-1} - \gamma \left[ f'_n(\bar{\theta}) + f''_n(\bar{\theta})(\theta_{n-1} - \bar{\theta}) \right]$ 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 - $p = 20$

![synthetic logistic – 1](image1)

![synthetic logistic – 2](image2)

\[
\log_{10}(n) \quad \log_{10}(f(\theta) - f(\theta^*))
\]

<table>
<thead>
<tr>
<th>1/2R^2</th>
<th>1/8R^2</th>
<th>1/32R^2</th>
<th>1/2R^2 n^{1/2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>green</td>
<td>red</td>
<td>cyan</td>
</tr>
</tbody>
</table>

Every iteration, every 2^p, 2-step, 2-step–dbl.
Simulations - benchmarks

- \textit{alpha} \( (p = 500, n = 500\,000) \), \textit{news} \( (p = 1\,300\,000, n = 20\,000) \)
Outline

1. Large-scale machine learning and optimization
   - Traditional statistical analysis
   - Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   - Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
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 $(x_1, y_1, \ldots, x_n, y_n)$
  - Multiple passes
  - Minimizes training cost $\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^\top \Phi(x_i))$
  - Need to regularize (e.g., by the $\ell_2$-norm) to avoid overfitting

• **Goal**: minimize $g(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$
Stochastic vs. deterministic methods

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

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

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

- **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 $\{1, \ldots, n\}$
  - 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 $f_i(\theta) = \ell(y_i, \theta^\top \Phi(x_i)) + \mu \Omega(\theta)$

- **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

\[
\log(\text{excess cost})
\]

\[
\text{time}
\]

- **stochastic**
- **deterministic**
Stochastic vs. deterministic methods

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

![Graph showing comparison between stochastic, deterministic, and hybrid methods](chart.png)
Accelerating gradient methods - Related work

- Nesterov acceleration
  - 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 \( f_i, i = 1, \ldots, n \)
  - Random selection \( i(t) \in \{1, \ldots, n\} \) 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 $f_i$, $i = 1, \ldots, n$
  - Random selection $i(t) \in \{1, \ldots, n\}$ with replacement
  - Iteration: $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^{n} y_t^i$ with $y_t^i = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_{i}^{t-1} & \text{otherwise} \end{cases}$

- 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 $L$-smooth, $i = 1, \ldots, n$
  - $g = \frac{1}{n} \sum_{i=1}^{n} f_i$ is $\mu$-strongly convex (with potentially $\mu = 0$)
  - constant step size $\gamma_t = 1/(16L)$
  - initialization with one pass of averaged SGD
Stochastic average gradient - Convergence analysis

• Assumptions
  – Each $f_i$ is $L$-smooth, $i = 1, \ldots, n$
  – $g = \frac{1}{n} \sum_{i=1}^{n} f_i$ is $\mu$-strongly convex (with potentially $\mu = 0$)
  – constant step size $\gamma_t = \frac{1}{16L}$
  – initialization with one pass of averaged SGD

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

$$
\mathbb{E}[g(\theta_t) - g(\theta^*)] \leq \left( \frac{8\sigma^2}{n\mu} + \frac{4L\|\theta_0 - \theta^*\|^2}{n} \right) \exp \left( -t \min \left\{ \frac{1}{8n}, \frac{\mu}{16L} \right\} \right)
$$

  – Linear (exponential) convergence rate with $O(1)$ iteration cost
  – After one pass, reduction of cost by $\exp \left( -\min \left\{ \frac{1}{8}, \frac{n\mu}{16L} \right\} \right)$
Stochastic average gradient - Convergence analysis

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

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

\[
\mathbb{E}[g(\theta_t) - g(\theta_\ast)] \leq 48\frac{\sigma^2 + L\|\theta_0 - \theta_\ast\|^2}{\sqrt{n}} \frac{n}{t}
\]

  – Improvement over regular batch and stochastic gradient
  – Adaptivity to potentially hidden strong convexity
**Convergence analysis - Proof sketch**

- **Main step:** find “good” Lyapunov function $J(\theta_t, y^t_1, \ldots, y^t_n)$
  - such that $\mathbb{E}[J(\theta_t, y^t_1, \ldots, y^t_n) | \mathcal{F}_{t-1}] < J(\theta_{t-1}, y^{t-1}_1, \ldots, y^{t-1}_n)$
  - no natural candidates

- **Computer-aided proof**
  - Parameterize function $J(\theta_t, y^t_1, \ldots, y^t_n) = g(\theta_t) - g(\theta_*) + \text{quadratic}$
  - Solve semidefinite program to obtain candidates (that depend on $n, \mu, L$)
  - 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 \textit{sparsity} in the features to reduce the storage and iteration cost

- \textbf{Grouping functions together} can further reduce the memory requirement

- We have obtained good performance when $L$ is not known with a \textit{heuristic line-search}

- Algorithm allows \textit{non-uniform sampling}

- Possibility of making \textit{proximal, coordinate-wise, and Newton-like variants}
spam dataset \( n = 92\,189, \ d = 823\,470 \)
Summary and future work

• Constant-step-size averaged stochastic gradient descent
  – Reaches convergence rate $O(1/n)$ in all regimes
  – Improves on the $O(1/\sqrt{n})$ lower-bound of non-smooth problems
  – Efficient online Newton step for non-quadratic problems
  – Robustness to step-size selection

• Going beyond a single pass through the data
Summary and future work

- **Constant-step-size averaged stochastic gradient descent**
  - Reaches convergence rate $O(1/n)$ in all regimes
  - Improves on the $O(1/\sqrt{n})$ lower-bound of non-smooth problems
  - Efficient online Newton step for non-quadratic problems
  - Robustness to step-size selection

- **Going beyond a single pass through the data**

- **Extensions and future work**
  - Pre-conditioning
  - Proximal extensions for non-differentiable terms
  - Kernels and non-parametric estimation
  - Line-search
  - Parallelization
Outline

1. Large-scale machine learning and optimization
   - Traditional statistical analysis
   - Classical methods for convex optimization

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

3. Smooth stochastic approximation algorithms
   - Asymptotic and non-asymptotic results

4. Beyond decaying step-sizes

5. Finite data sets
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
  – Going beyond a single pass over the data (testing performance)
  – Characterization of implicit regularization of online methods
  – Further links between convex optimization and online learning/bandits
References


S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic


O. Macchi. *Adaptive processing: The least mean squares approach with applications in transmission*. 


P. Sunehag, J. Trumpf, SVN Vishwanathan, and N. Schraudolph. Variable metric stochastic


