Machine learning challenges for big data

Francis Bach

SIERRA Project-team, INRIA - Ecole Normale Supérieure

Machine learning
Computer science and applied mathematics

• Modelisation, prediction and control from training examples

• Theory
  – Analysis of statistical performance

• Algorithms
  – Numerical efficiency and stability

• Applications
  – Computer vision, bioinformatics, neuro-imaging, text, audio
Context

Machine learning for “big data”

- **Large-scale machine learning**: large $p$, large $n$, large $k$
  - $p$: dimension of each observation (input)
  - $k$: number of tasks (dimension of outputs)
  - $n$: number of observations

- **Examples**: computer vision, bioinformatics, etc.
Context

Machine learning for “big data”

• **Large-scale machine learning:** large \( p \), large \( n \), large \( k \)
  
  – \( p \): dimension of each observation (input)
  – \( k \): number of tasks (dimension of outputs)
  – \( n \): number of observations

• **Examples:** computer vision, bioinformatics, etc.

• Two main challenges:

  1. **Computational:** ideal running-time complexity \( = O(pn + kn) \)
  2. **Statistical:** meaningful results
Machine learning challenges for big data

Recent work

1. **Large-scale** supervised learning
   - Going beyond stochastic gradient descent
   - Le Roux, Schmidt, and Bach (2012)

2. **Unsupervised** learning through dictionary learning
   - Imposing structure for interpretability
   - Bach, Jenatton, Mairal, and Obozinski (2011, 2012)

3. **Interactions between convex and combinatorial** optimization
   - Submodular functions
   - Bach (2011); Obozinski and Bach (2012)
Supervised machine learning

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

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

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

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

convex data fitting term + regularizer
Supervised machine learning

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

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

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

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

convex data fitting term + regularizer

- Applications to any data-oriented field
  - Computer vision, bioinformatics
  - Natural language processing, etc.
Supervised machine learning

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

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

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

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

convex data fitting term + regularizer

- **Main practical challenges**
  - Designing/learning good features \( \Phi(x) \)
  - Efficiently solving the optimization problem
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
  – Iteration complexity is linear in $n$
Stochastic vs. deterministic methods

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

- 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
  - Iteration complexity is linear in $n$

- **Stochastic** gradient descent: $\theta_t = \theta_{t-1} - \gamma_t f_{i(t)}'(\theta_{t-1})$
  
  - Sampling with replacement: $i(t)$ random element of $\{1, \ldots, n\}$
  - Convergence rate in $O(1/t)$
  - Iteration complexity is independent of $n$
Stochastic vs. deterministic methods

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

- **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: $\theta_t = \theta_{t-1} - \gamma_t f_i'(\theta_{t-1})$
Stochastic vs. deterministic methods

- **Goal** = best of both worlds: linear rate with $O(1)$ iteration cost
Stochastic vs. deterministic methods

- **Goal** = best of both worlds: linear rate with $O(1)$ iteration cost
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 \( y_i^t = \begin{cases} f'_i(\theta_{t-1}) & \text{if } i = i(t) \\ y_{i-1}^t & \text{otherwise} \end{cases} \)

- Stochastic version of incremental average gradient (Blatt et al., 2008)

- Simple implementation
  - Extra memory requirement: same size as original data (or less)
Stochastic average gradient

Convergence analysis

- Assume each $f_i$ is $L$-smooth and $g = \frac{1}{n} \sum_{i=1}^{n} f_i$ is $\mu$-strongly convex

- Constant step size $\gamma_t = \frac{1}{2n\mu}$. If $\frac{\mu}{L} \geq \frac{8}{n}$, then $\exists C \in \mathbb{R}$ such that

\[
\forall t \geq 0, \quad \mathbb{E}[g(\theta_t) - g(\theta^*)] \leq C \left(1 - \frac{1}{8n}\right)^t
\]
Stochastic average gradient
Convergence analysis

• Assume each $f_i$ is $L$-smooth and $g = \frac{1}{n} \sum_{i=1}^{n} f_i$ is $\mu$-strongly convex

• Constant step size $\gamma_t = \frac{1}{2n\mu}$. If $\frac{\mu}{L} \geq \frac{8}{n}$, then $\exists C \in \mathbb{R}$ such that

$$\forall t \geq 0, \quad \mathbb{E}[g(\theta_t) - g(\theta^*)] \leq C \left(1 - \frac{1}{8n}\right)^t$$

• Linear convergence rate with iteration cost independent of $n$

• Linear convergence rate “independent” of the condition number
  – After each pass through the data, constant error reduction
  – Application to linear systems
Stochastic average gradient
Simulation experiments

• protein dataset (n = 145751, p = 74)

• Dataset split in two (training/testing)

Training cost

Testing cost
Stochastic average gradient

Simulation experiments

- cover type dataset \((n = 581012, p = 54)\)

- Dataset split in two (training/testing)
Machine learning challenges for big data

Recent work

1. **Large-scale supervised learning**
   - Going beyond stochastic gradient descent
   - Le Roux, Schmidt, and Bach (2012)

2. **Unsupervised learning through dictionary learning**
   - Imposing structure for interpretability
   - Bach, Jenatton, Mairal, and Obozinski (2011, 2012)

3. **Interactions between convex and combinatorial optimization**
   - Submodular functions
   - Bach (2011); Obozinski and Bach (2012)
Learning dictionaries for uncovering hidden structure

- **Fact**: many natural signals may be approximately represented as a superposition of few atoms from a dictionary $D = (d_1, \ldots, d_p)$

  - Decomposition $x = \sum_{i=1}^{p} \alpha_i d_i$ with $\alpha \in \mathbb{R}^p$ sparse

  - Natural signals (sounds, images) and others

- **Decoding problem**: given a dictionary $D$, finding $\alpha$ through regularized convex optimization $\min_{\alpha \in \mathbb{R}^p} \| x - \sum_{i=1}^{p} \alpha_i d_i \|_2^2 + \lambda \| \alpha \|_1$
Learning dictionaries for uncovering hidden structure

- **Fact:** many natural signals may be approximately represented as a superposition of few atoms from a dictionary $D = (d_1, \ldots, d_p)$
  
  - Decomposition $x = \sum_{i=1}^{p} \alpha_i d_i$ with $\alpha \in \mathbb{R}^p$ sparse
  
  - Natural signals (sounds, images) and others

- **Decoding problem:** given a dictionary $D$, finding $\alpha$ through regularized convex optimization $\min_{\alpha \in \mathbb{R}^p} \|x - \sum_{i=1}^{p} \alpha_i d_i\|_2^2 + \lambda \|\alpha\|_1$

- **Dictionary learning problem:** given $n$ signals $x^1, \ldots, x^n$,
  
  - Estimate both dictionary $D$ and codes $\alpha^1, \ldots, \alpha^n$

$$\min_D \sum_{j=1}^{n} \min_{\alpha^j \in \mathbb{R}^p} \left\{ \|x^j - \sum_{i=1}^{p} \alpha^j_i d_i\|_2^2 + \lambda \|\alpha^j\|_1 \right\}$$
Challenges of dictionary learning

\[
\min_D \sum_{j=1}^{n} \min_{\alpha^j \in \mathbb{R}^p} \left\{ \| x^j - \sum_{i=1}^{p} \alpha^j_i d_i \|_2^2 + \lambda \| \alpha^j \|_1 \right\}
\]

- **Algorithmic challenges**
  - Large number of signals \( \Rightarrow \) online learning (Mairal et al., 2009)

- **Theoretical challenges**
  - Identifiability/robustness (Jenatton et al., 2012)

- **Domain-specific challenges**
  - Going beyond plain sparsity \( \Rightarrow \) structured sparsity
    (Jenatton, Mairal, Obozinski, and Bach, 2011)
Structured sparsity

- Sparsity-inducing behavior depends on “corners” of unit balls
Structured sparse PCA (Jenatton et al., 2009)

- Unstructured sparse PCA \( \Rightarrow \) many zeros do not lead to better interpretability
Structured sparse PCA (Jenatton et al., 2009)

- Unstructured sparse PCA → many zeros do not lead to better interpretability
Structured sparse PCA (Jenatton et al., 2009)

- Enforce selection of **convex** nonzero patterns $\Rightarrow$ robustness to occlusion in face identification
Structured sparse PCA (Jenatton et al., 2009)

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion in face identification
Recent work

1. **Large-scale supervised learning**
   - Going beyond stochastic gradient descent
   - Le Roux, Schmidt, and Bach (2012)

2. **Unsupervised learning through dictionary learning**
   - Imposing structure for interpretability
   - Bach, Jenatton, Mairal, and Obozinski (2011, 2012)

3. **Interactions between convex and combinatorial optimization**
   - Submodular functions
   - Bach (2011); Obozinski and Bach (2012)
Conclusion and open problems
Machine learning for “big data”

- Having a large-scale hardware infrastructure is not enough

- Large-scale learning
  - Between theory, algorithms and applications
  - Adaptivity to increased amounts of data with linear complexity
  - Robust algorithms with no hyperparameters

- Unsupervised learning
  - Incorporating structural prior knowledge
  - Semi-supervised learning
  - Automatic learning of features for supervised learning
References

http://hal.inria.fr/hal-00645271/en.


D. Blatt, A.O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical


N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence

International Conference on Machine Learning (ICML), 2009.