Machine learning challenges for big data

Francis Bach

SIERRA Project-team, INRIA - Ecole Normale Supérieure

Joint work with R. Jenatton, J. Mairal, G. Obozinski, N. Le Roux, M. Schmidt - December 2012

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 $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- 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

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

convex data fitting term + regularizer

Supervised machine learning

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

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

convex data fitting term $+$ regularizer

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

Supervised machine learning

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

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

convex data fitting term + regularizer

- Main practical challenges
  - Designing/learning good features $\Phi(x)$
  - Efficiently solving the optimization problem

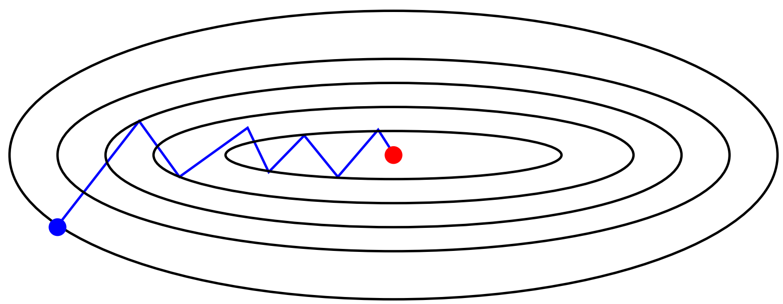
Stochastic vs. deterministic methods

- Minimizing $g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ with $f_i(\theta) = \ell\big(y_i, \theta^\top \Phi(x_i)\big) + \mu\,\Omega(\theta)$
- Batch gradient descent: $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n}\sum_{i=1}^{n} f_i'(\theta_{t-1})$
  - Linear (e.g., exponential) convergence rate
  - Iteration complexity is linear in n

Stochastic vs. deterministic methods

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

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



Stochastic vs. deterministic methods

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

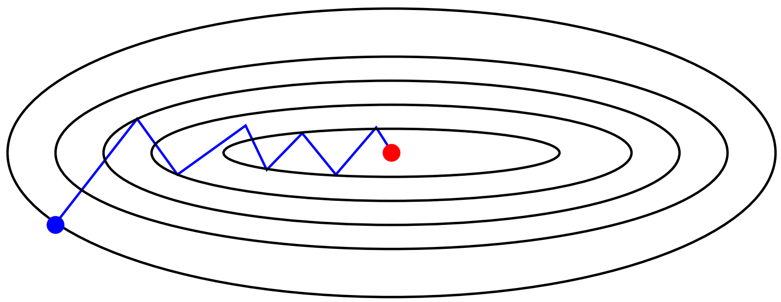
- Batch gradient descent: $\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1}) = \theta_{t-1} - \frac{\gamma_t}{n}\sum_{i=1}^{n} f_i'(\theta_{t-1})$
  - Linear (e.g., exponential) convergence rate
  - Iteration complexity is linear in n

- Stochastic gradient descent: $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$
  - Sampling with replacement: $i(t)$ random element of $\{1, \dots, n\}$
  - Convergence rate in $O(1/t)$
  - Iteration complexity is independent of n

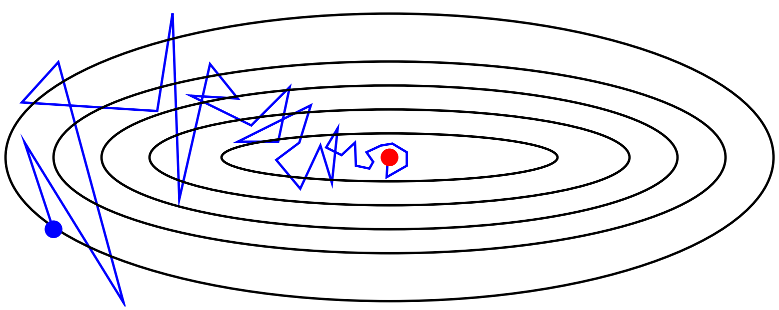
Stochastic vs. deterministic methods

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

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

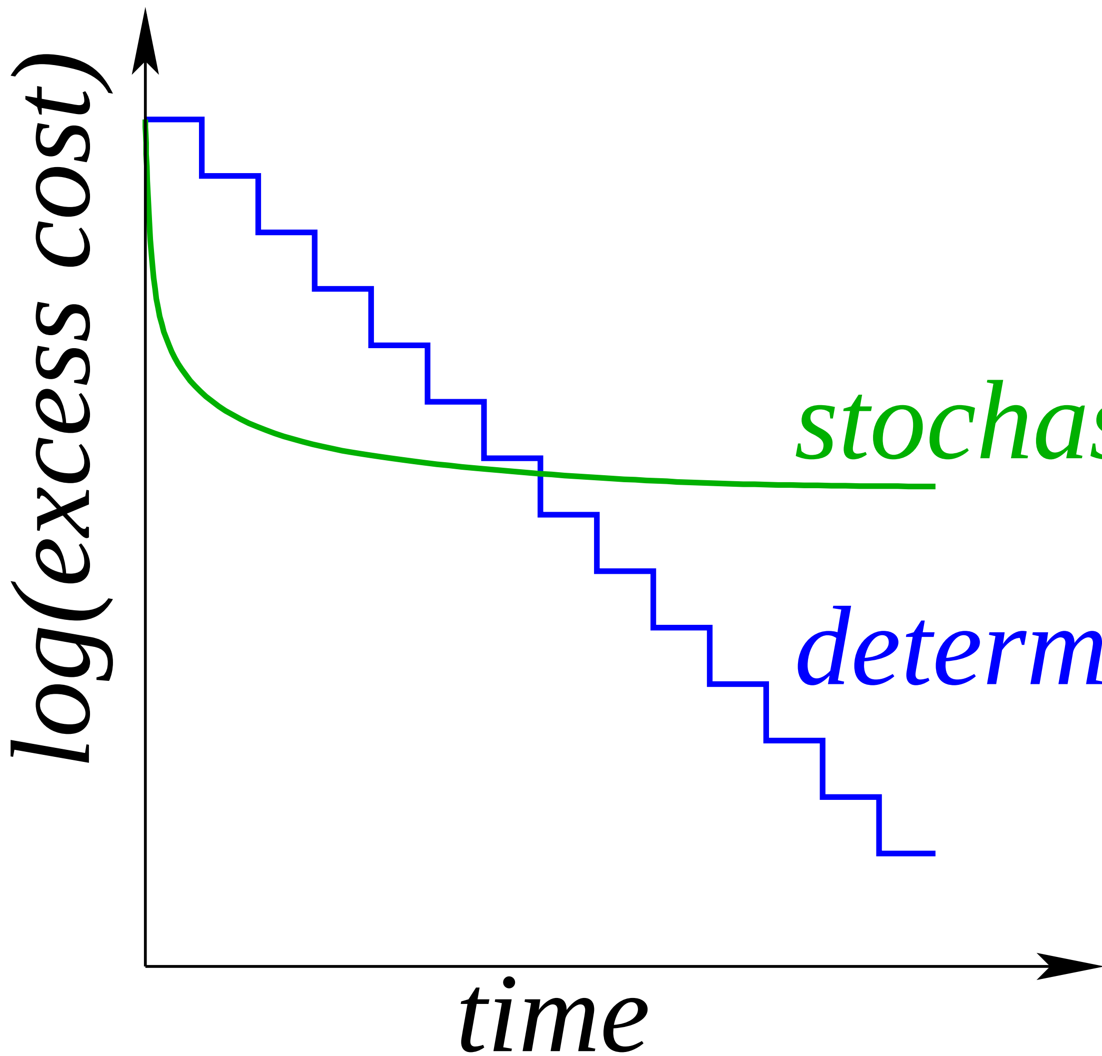


- Stochastic gradient descent: $\theta_t = \theta_{t-1} - \gamma_t f'_{i(t)}(\theta_{t-1})$



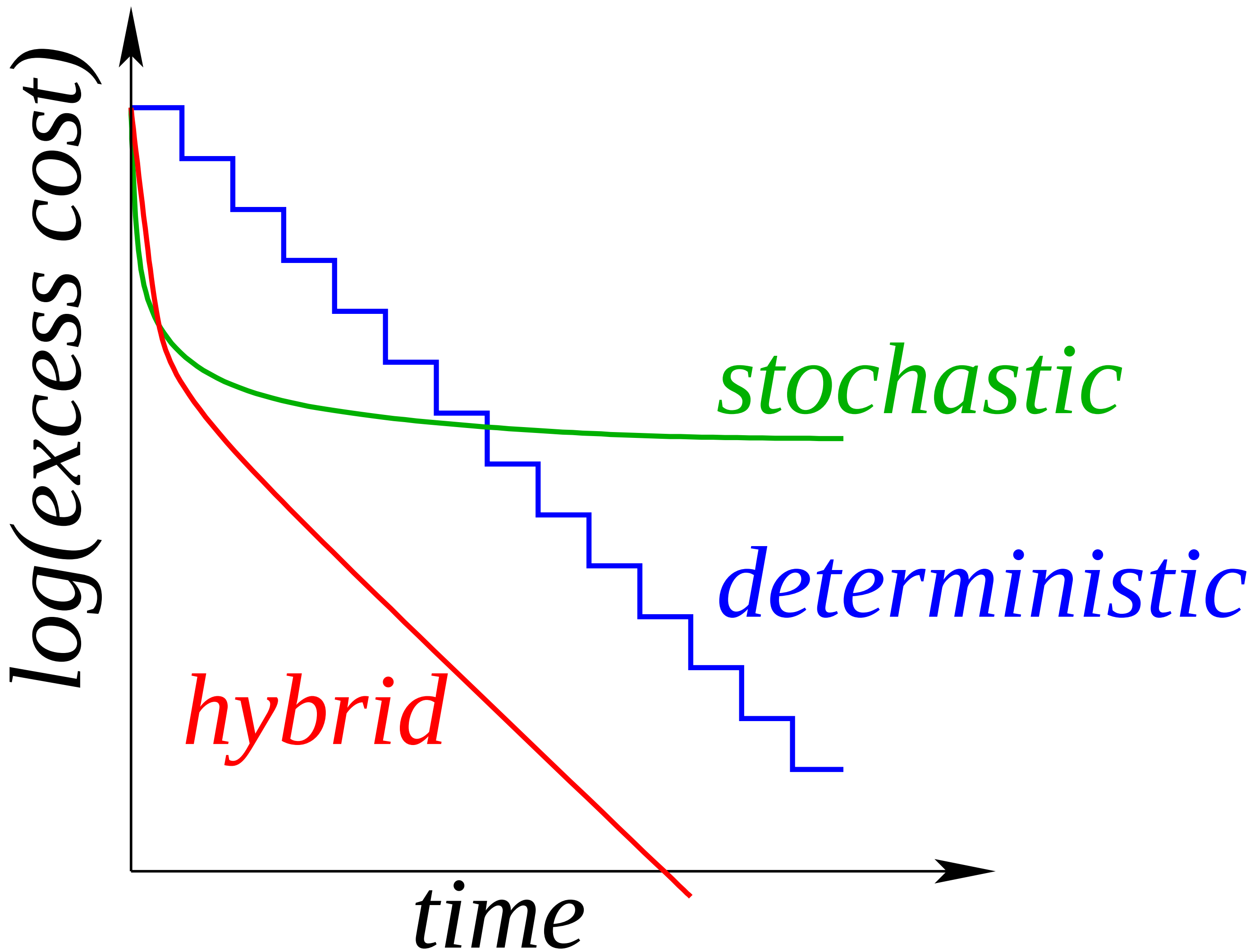
Stochastic vs. deterministic methods

- **Goal** = **best of both worlds**: linear rate with $O(1)$ iteration cost



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, $i = 1, \dots, n$
  - Random selection $i(t) \in \{1, \dots, n\}$ with replacement
  - Iteration: $\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^{n} y_i^t$ with $y_i^t = \begin{cases} f_i'(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{otherwise} \end{cases}$
- Stochastic version of incremental average gradient (Blatt et al., 2008)
- 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 μ-strongly convex

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

$$\forall t \geqslant 0,\ \mathbb{E}\big[g(\theta_t) - g(\theta^*)\big] \leqslant C\Big(1 - \frac{1}{8n}\Big)^t$$

Stochastic average gradient
Convergence analysis

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

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

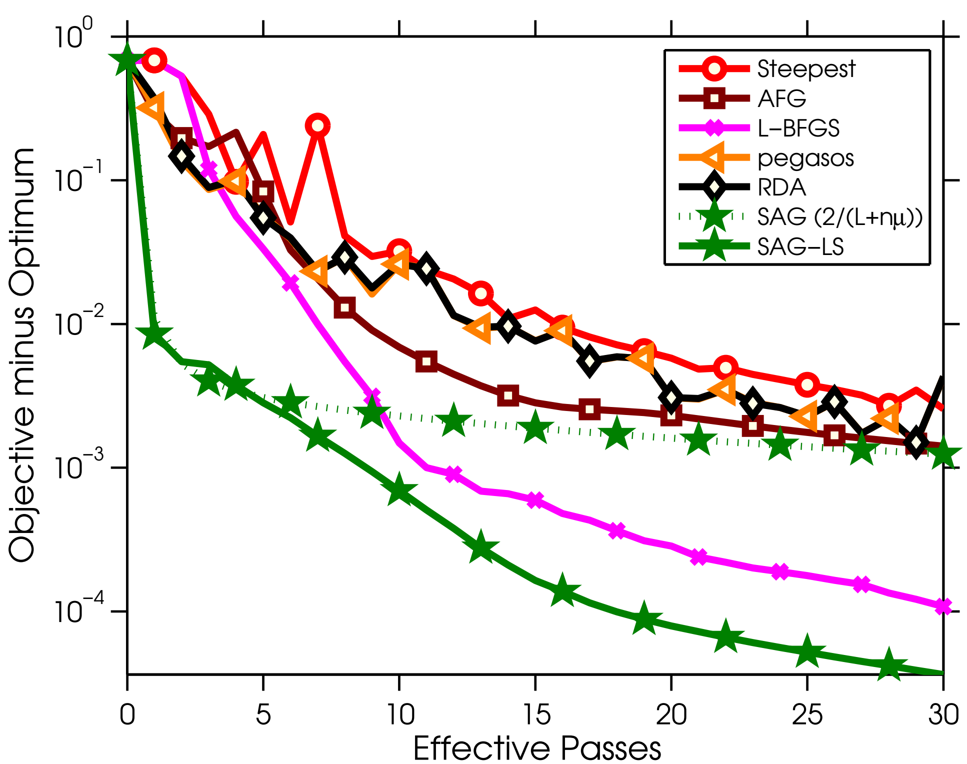
$$\forall t \geqslant 0, \ \mathbb{E}\big[g(\theta_t) - g(\theta^*)\big] \leqslant C\Big(1 - \frac{1}{8n}\Big)^t$$

- **Linear convergence rate with iteration cost independent of n**

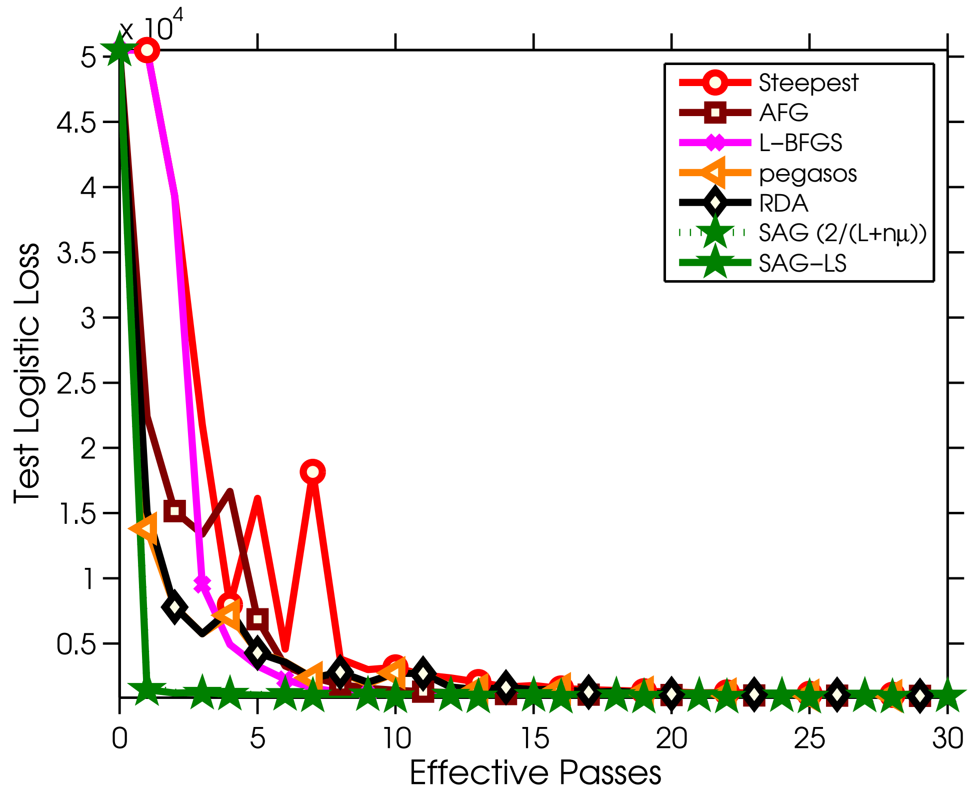
- Linear convergence rate "independent" of the condition number
  - After each pass through the data, constant error reduction
  - Application to linear systems

Stochastic average gradient
Simulation experiments

- `protein` dataset ($n = 145751$, $p = 74$)
- Dataset split in two (training/testing)



Training cost

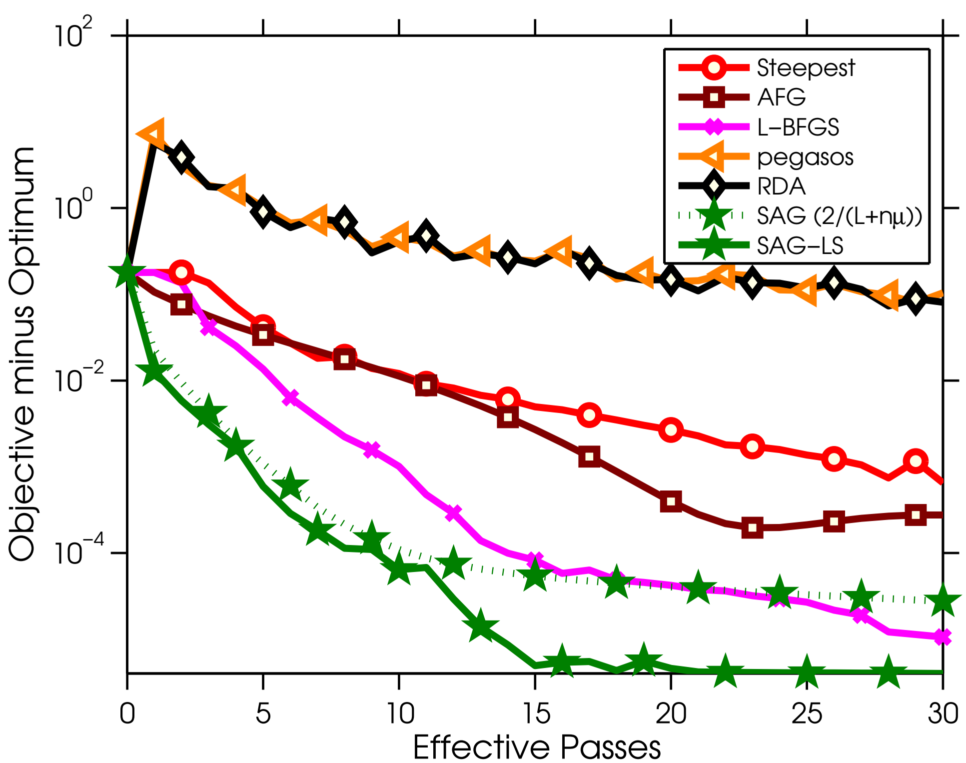


Testing cost

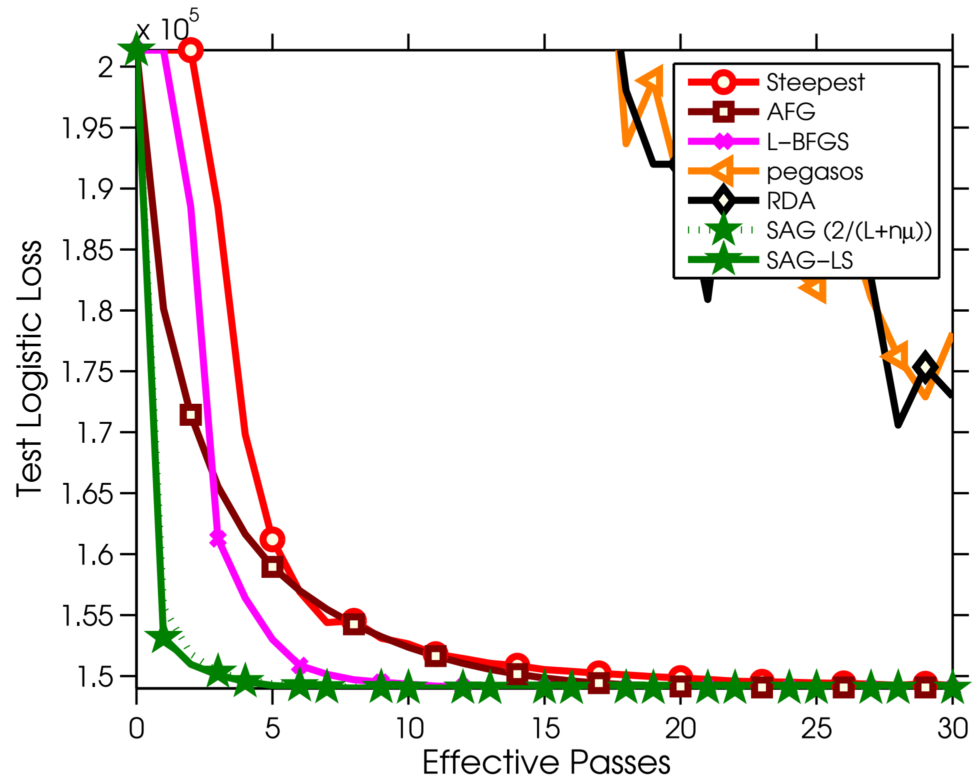
Stochastic average gradient

Simulation experiments

- `cover type` dataset ($n = 581012$, $p = 54$)
- Dataset split in two (training/testing)



Training cost



Testing cost

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 $D = (d_1, \dots, d_p)$
  - 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 α 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 $D = (d_1, \dots, d_p)$
  - 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 α 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 $x^1, \dots, x^n$,
  - Estimate both dictionary D and codes $\alpha^1, \dots, \alpha^n$

$$\min_{D} \sum_{j=1}^{n} \min_{\alpha^j \in \mathbb{R}^p} \left\{ \left\| x^j - \sum_{i=1}^{p} \alpha_i^j d_i \right\|_2^2 + \lambda \|\alpha^j\|_1 \right\}$$

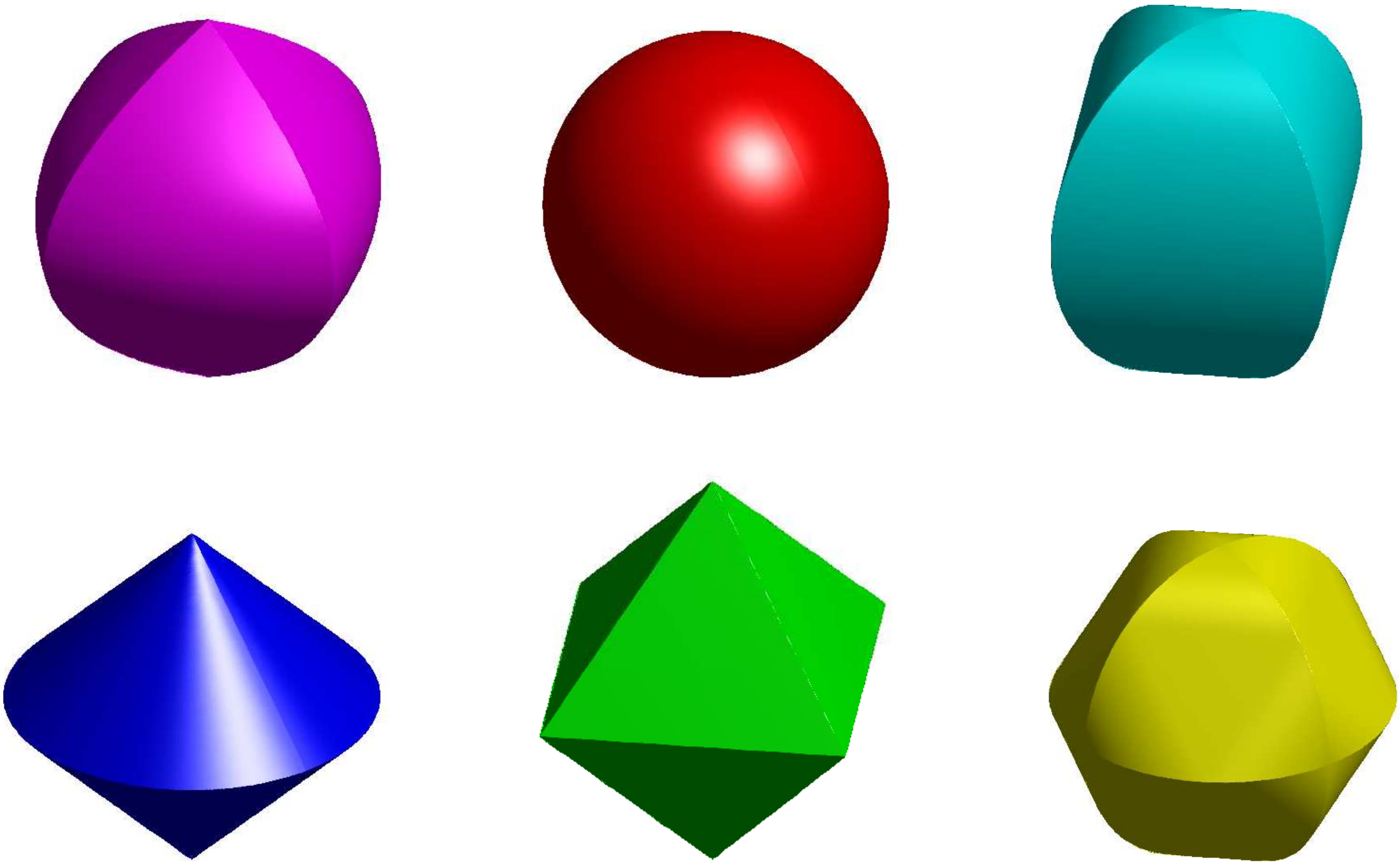
Challenges of dictionary learning

$$\min_{D} \sum_{j=1}^{n} \min_{\alpha^j \in \mathbb{R}^p} \left\{ \left\| x^j - \sum_{i=1}^{p} \alpha_i^j d_i \right\|_2^2 + \lambda \|\alpha^j\|_1 \right\}$$

- **Algorithmic challenges**
  - Large number of signals $\Rightarrow$ online learning (Mairal et al., 2009)
- **Theoretical challenges**
  - Identifiabiliy/robustness (Jenatton et al., 2012)
- **Domain-specific challenges**
  - Going beyond plain sparsity $\Rightarrow$ structured sparsity (Jenatton, Mairal, Obozinski, and Bach, 2011)

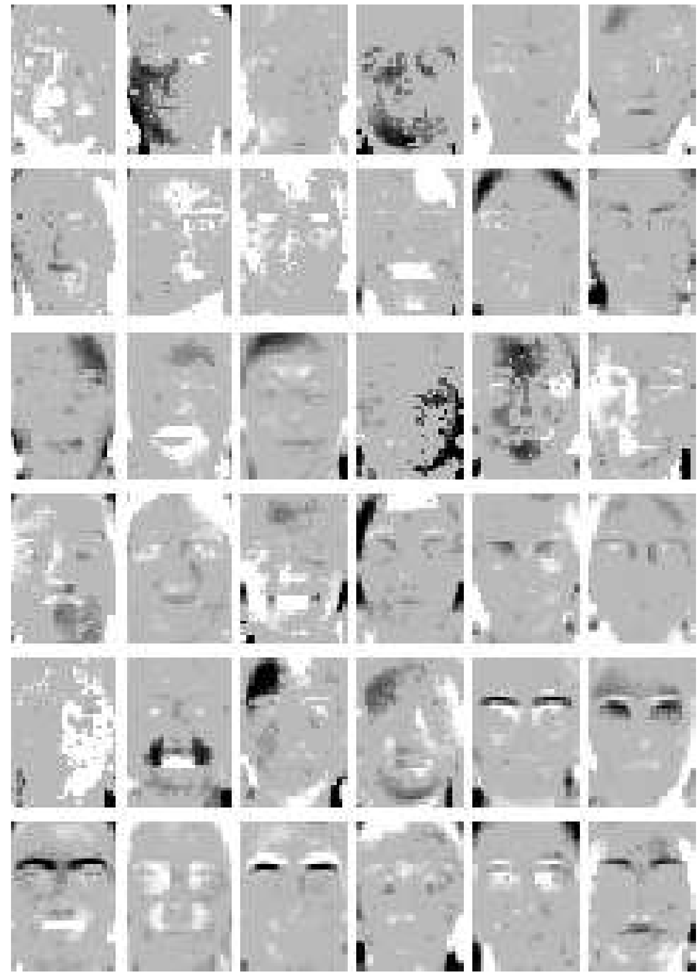
Structured sparsity

- **Sparsity-inducing behavior depends on "corners" of unit balls**



Structured sparse PCA (Jenatton et al., 2009)

raw data

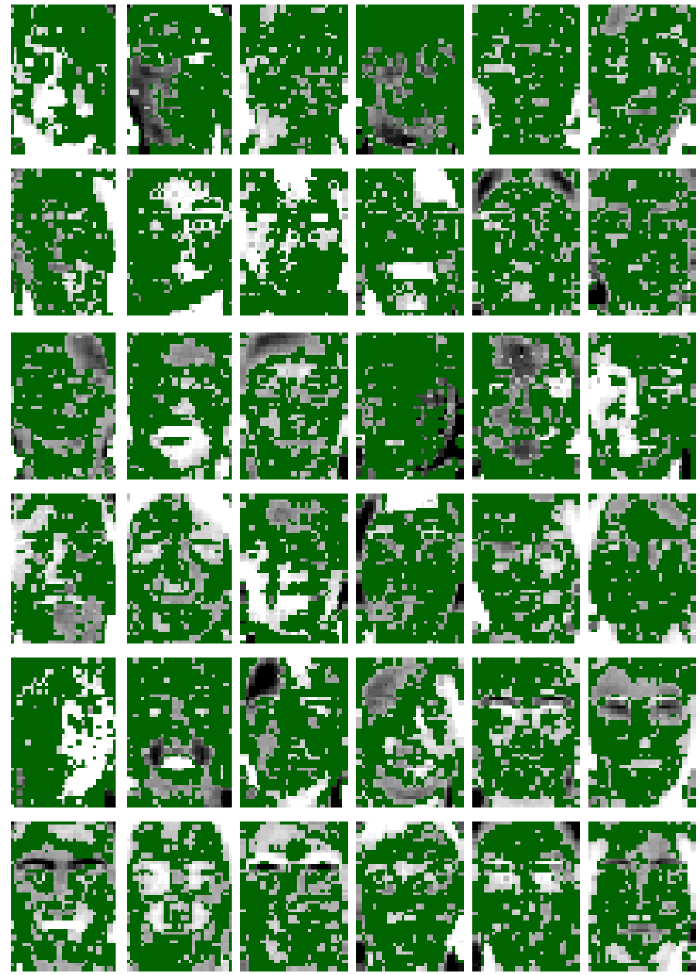


sparse PCA

- Unstructured sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009)

raw data

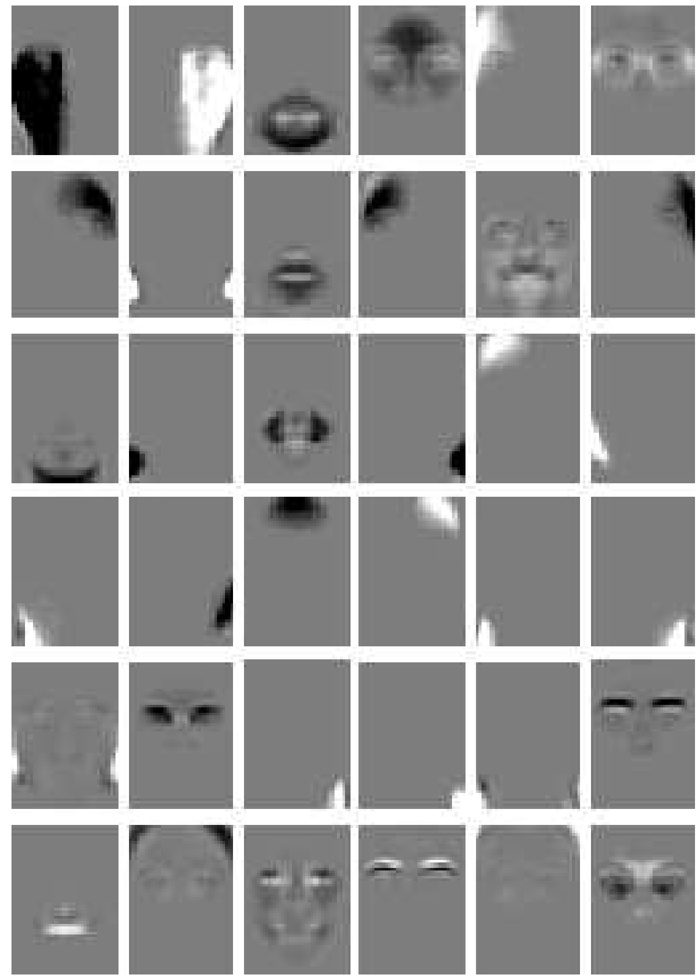


sparse PCA

- Unstructured sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009)

raw data

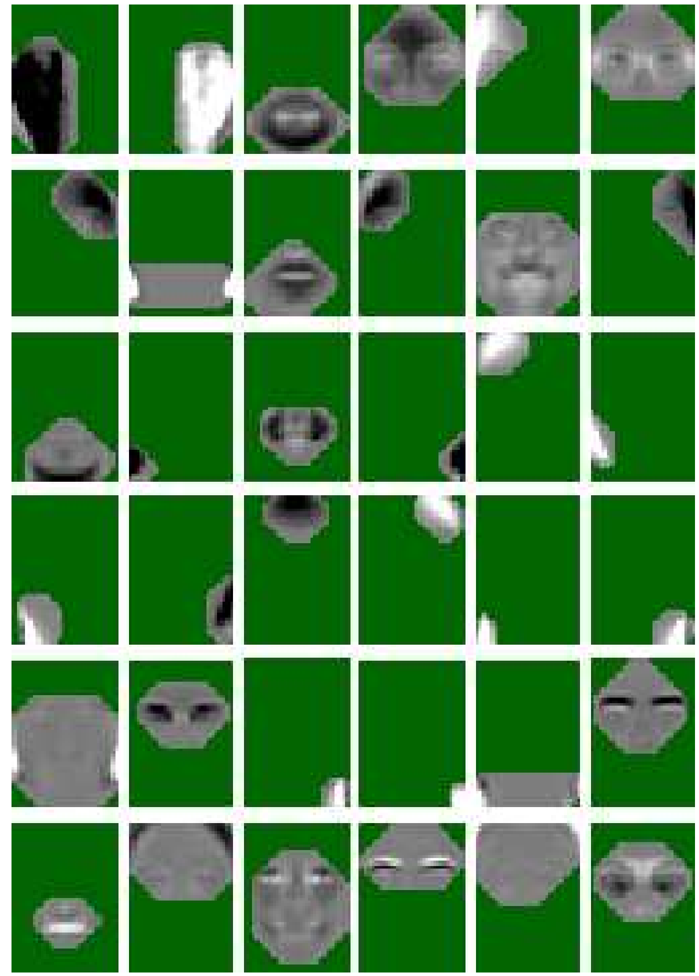


Structured sparse PCA

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

Structured sparse PCA (Jenatton et al., 2009)

raw data



Structured sparse PCA

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

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)

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

F. Bach. Learning with Submodular Functions: A Convex Optimization Perspective. 2011. URL `http://hal.inria.fr/hal-00645271/en`.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2011.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 2012. To appear.

D. Blatt, A.O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. 18(1):29–51, 2008.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. Technical Report -, HAL, 2012.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*, 2009.

G. Obozinski and F. Bach. Convex relaxation of combinatorial penalties. Technical report, HAL, 2012.