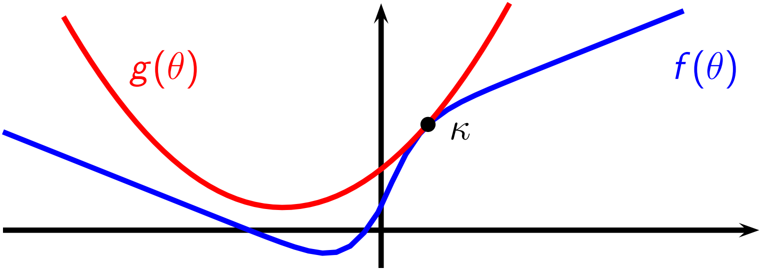
Incremental and Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization

Julien Mairal

INRIA LEAR, Grenoble

Gargantua workshop, LJK, November 2013

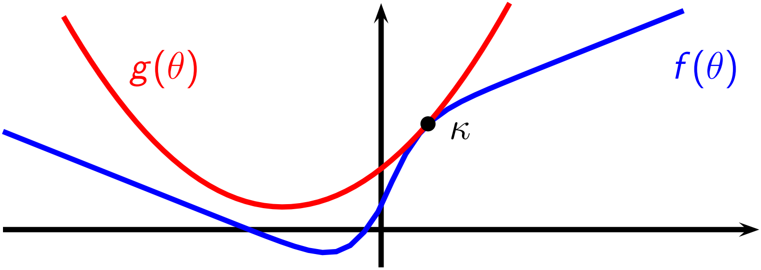
A simple optimization principle



Objective: $\min_{\theta \in \Theta} f(\theta)$

- Principle called Majorization-Minimization [Lange et al., 2000];
- quite popular in statistics and signal processing.

In this work

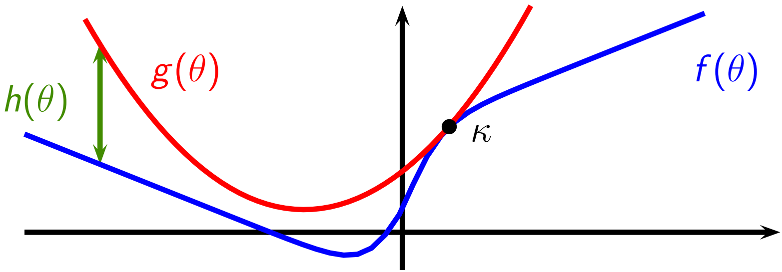


- **scalable** Majorization-Minimization algorithms;
- for **convex or non-convex** and **smooth or non-smooth** problems;

References

- J. Mairal. Optimization with First-Order Surrogate Functions. ICML'13;
- J. Mairal. Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization. NIPS'13.

Setting: First-Order Surrogate Functions



- $g(\theta') \geq f(\theta')$ for all θ' in $\arg\min_{\theta \in \Theta} g(\theta)$;
- the **approximation error** $h \stackrel{\triangle}{=} g - f$ is differentiable, and ∇h is L-Lipschitz. Moreover, $h(\kappa) = 0$ and $\nabla h(\kappa) = 0$.

The Basic MM Algorithm

Algorithm 1 Basic Majorization-Minimization Scheme

Input: $\theta_0 \in \Theta$ (initial estimate); N (number of iterations).
for $n = 1, \ldots, N$ **do**
  Compute a surrogate g_n of f near θ_{n-1};
  Minimize g_n and update the solution:

$$\theta_n \in \arg\min_{\theta \in \Theta} g_n(\theta).$$

end for
Output: θ_N (final estimate);

Examples of First-Order Surrogate Functions

- **Lipschitz Gradient Surrogates**:
  f is L-smooth (differentiable + L-Lipschitz gradient).

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top (\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2.$$

  Minimizing g yields a gradient descent step $\theta \leftarrow \kappa - \frac{1}{L}\nabla f(\kappa)$.

Examples of First-Order Surrogate Functions

- **Lipschitz Gradient Surrogates**:
  f is L-smooth (differentiable + L-Lipschitz gradient).

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top(\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2.$$

  Minimizing g yields a gradient descent step $\theta \leftarrow \kappa - \frac{1}{L}\nabla f(\kappa)$.

- **Proximal Gradient Surrogates**:
  $f = f_1 + f_2$ with f_1 smooth.

$$g : \theta \mapsto f_1(\kappa) + \nabla f_1(\kappa)^\top(\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2 + f_2(\theta).$$

  Minimizing g amounts to one step of the forward-backward, ISTA, or proximal gradient descent algorithm.
  [Beck and Teboulle, 2009, Combettes and Pesquet, 2010, Wright et al., 2008, Nesterov, 2007].

Examples of First-Order Surrogate Functions

- **Linearizing Concave Functions and DC-Programming**:
  $f = f_1 + f_2$ with f_2 smooth and concave.

$$g : \theta \mapsto f_1(\theta) + f_2(\kappa) + \nabla f_2(\kappa)^\top (\theta - \kappa).$$

When f_1 is convex, the algorithm is called DC-programming.

Examples of First-Order Surrogate Functions

- **Linearizing Concave Functions and DC-Programming**:
  $f = f_1 + f_2$ with f_2 smooth and concave.

$$g : \theta \mapsto f_1(\theta) + f_2(\kappa) + \nabla f_2(\kappa)^\top(\theta - \kappa).$$

  When f_1 is convex, the algorithm is called DC-programming.

- **Quadratic Surrogates**:
  f is twice differentiable, and $\mathbf{H}$ is a uniform upper bound of $\nabla^2 f$:

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top(\theta - \kappa) + \frac{1}{2}(\theta - \kappa)^\top \mathbf{H}(\theta - \kappa).$$

  Actually a big deal in statistics and machine learning [Böhning and Lindsay, 1988, Khan et al., 2010, Jebara and Choromanska, 2012].

Examples of First-Order Surrogate Functions

- **More Exotic Surrogates**:
  Consider a smooth approximation of the trace (nuclear) norm

$$f_\mu : \theta \mapsto \text{Tr}\left((\theta^\top \theta + \mu \mathbf{I})^{1/2}\right) = \sum_{i=1}^{p} \sqrt{\lambda_i(\theta^\top \theta) + \mu},$$

  $f' : \mathbf{H} \mapsto \text{Tr}\left(\mathbf{H}^{1/2}\right)$ is concave on the set of p.d. matrices and $\nabla f'(\mathbf{H}) = (1/2)\mathbf{H}^{-1/2}$.

$$g_\mu : \theta \mapsto f_\mu(\kappa) + \frac{1}{2}\text{Tr}\left((\kappa^\top \kappa + \mu \mathbf{I})^{-1/2}(\theta^\top \theta - \kappa^\top \kappa)\right),$$

  which yields the algorithm of Mohan and Fazel [2012].

Examples of First-Order Surrogate Functions

- **Variational Surrogates**: $f(\theta_1) \triangleq \min_{\theta_2 \in \Theta_2} \tilde{f}(\theta_1, \theta_2)$,
  where $\tilde{f}$ is "smooth" w.r.t θ_1 and strongly convex w.r.t θ_2:

$$g : \theta_1 \mapsto \tilde{f}(\theta_1, \kappa_2^\star) \text{ with } \kappa_2^\star \triangleq \operatorname*{arg\,min}_{\theta_2 \in \Theta_2} \tilde{f}(\kappa_1, \theta_2).$$

- **Saddle-Point Surrogates**: $f(\theta_1) \triangleq \max_{\theta_2 \in \Theta_2} \tilde{f}(\theta_1, \theta_2)$,
  where $\tilde{f}$ is "smooth" w.r.t θ_1 and strongly concave w.r.t θ_2:

$$g : \theta_1 \mapsto \tilde{f}(\theta_1, \kappa_2^\star) + \frac{L''}{2}\|\theta_1 - \kappa_1\|_2^2.$$

- **Jensen Surrogates**: $f(\theta) \triangleq \tilde{f}(\mathbf{x}^\top \theta)$,
  where $\tilde{f}$ is L-smooth. Choose a weight vector $\mathbf{w}$ in $\mathbb{R}_+^p$ such that $\|\mathbf{w}\|_1 = 1$ and $\mathbf{w}_i \neq 0$ whenever $\mathbf{x}_i \neq 0$.

$$g : \theta \mapsto \sum_{i=1}^{p} \mathbf{w}_i f\left(\frac{\mathbf{x}_i}{\mathbf{w}_i}(\theta_i - \kappa_i) + \mathbf{x}^\top \kappa\right),$$

Theoretical Guarantees

- for **non-convex** problems: $f(\theta_n)$ monotically decreases and

$$\liminf_{n \to +\infty} \inf_{\theta \in \Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq 0,$$

  which is an asymptotic stationary point condition.
- for **convex** ones: $f(\theta_n) - f^\star = O(1/n)$.
- for μ-**strongly convex** ones: the convergence rate is linear $O((1 - \mu/L)^n)$.

the convergence rates and the proof techniques are the same as for proximal gradient methods [Nesterov, 2007, Beck and Teboulle, 2009].

New Majorization-Minimization Algorithms

Given $f : \mathbb{R}^p \to \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^p$, our goal is to solve

$$\min_{\theta \in \Theta} f(\theta).$$

We introduce algorithms for **non-convex and convex** optimization:

- a block coordinate scheme for separable surrogates;
- an **incremental** algorithm dubbed MISO for separable functions f;
- a **stochastic** algorithm for minimizing expectations;

Also several variants for **convex optimization**:

- an accelerated one (Nesterov's like);
- a "Frank-Wolfe" majorization-minimization algorithm.

Incremental Optimization: MISO

Suppose that f splits into many components:

$$f(\theta) = \frac{1}{T}\sum_{t=1}^{T} f^t(\theta).$$

Recipe

- Incrementally update an approximate surrogate $\frac{1}{T}\sum_{t=1}^{T} g^t$;
- add some heuristics for practical implementations.

Related (Inspiring) Work for Convex Problems

- related to SAG [Schmidt et al., 2013] and SDCA [Shalev-Schwartz and Zhang, 2012], but offers different update rules.

Incremental Optimization: MISO

Algorithm 2 Incremental Scheme MISO

Input: $\theta_0 \in \Theta$; N (number of iterations).
Choose surrogates g_0^t of f^t near θ_0 for all t;
for $n = 1, \ldots, N$ **do**
    Randomly pick up one index $\hat{t}_n$ and choose a surrogate $g_n^{\hat{t}_n}$ of $f^{\hat{t}_n}$ near θ_{n-1}. Set $g_n^t \triangleq g_{n-1}^t$ for $t \neq \hat{t}_n$;
    Update the solution:

$$\theta_n \in \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} g_n^t(\theta)$$

.
end for
Output: θ_N (final estimate);

Incremental Optimization: MISO

Update Rule for Proximal Gradient Surrogates

We want to minimize $\frac{1}{T}\sum_{t=1}^{T} f_1^t(\theta) + f_2(\theta)$.

$$\theta_n = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} f_1(\kappa^t) + \nabla f_1(\kappa^t)^\top(\theta - \kappa^t) + \frac{L}{2}\|\theta - \kappa^t\|_2^2 + f_2(\theta)$$
$$= \arg\min_{\theta\in\Theta} \frac{1}{2}\left\|\theta - \left(\frac{1}{T}\sum_{t=1}^{T}\kappa^t - \frac{1}{LT}\sum_{t=1}^{T}\nabla f_1^t(\kappa^t)\right)\right\|_2^2 + \frac{1}{L}f_2(\theta).$$

Then, randomly draw one index t_n, and update $\kappa^{t_n} \leftarrow \theta_n$.

Remark

- remove f_2, replace $\frac{1}{T}\sum_{t=1}^{T}\kappa^t$ by θ_{n-1} yields SAG [Schmidt et al., 2013];
- replace L by μ is "close" to SDCA [Shalev-Schwartz and Zhang, 2012];

Incremental Optimization: MISO

Theoretical Guarantees

- for **non-convex** problems, the guarantees are the same as the generic MM algorithm with probability one.
- for **convex** problems and proximal gradient surrogates, the expected convergence rate becomes $O(T/n)$.
- for μ-**strongly convex** problems and proximal gradient surrogates, the expected convergence rate is linear $O((1-\mu/(TL))^n)$.

Incremental Optimization: MISO

Theoretical Guarantees

- for **non-convex** problems, the guarantees are the same as the generic MM algorithm with probability one.
- for **convex** problems and proximal gradient surrogates, the expected convergence rate becomes $O(T/n)$.
- for μ-**strongly convex** problems and proximal gradient surrogates, the expected convergence rate is linear $O((1-\mu/(TL))^n)$.

Remarks

- for μ-strongly convex problems, the rates of SDCA and SAG are better: $\mu/(LT)$ is replaced by $O(\min(\mu/L, 1/T))$;
- MISO with minorizing surrogates is close to SDCA with "similar" convergence rates (details to be written yet).

Stochastic Majorization Minimization: SMM

Suppose that f is an expectation:

$$f(\theta) = \mathbb{E}_{\mathbf{x}}[l(\theta, \mathbf{x})].$$

Recipe

- Draw a function $f_n : \theta \mapsto l(\theta, \mathbf{x}_n)$ at iteration n;
- Iteratively update an approximate surrogate $\bar{g}_n = (1 - w_n)\bar{g}_{n-1} + w_n g_n$;
- Possibly use an averaging scheme of the iterates.

Related Work

- online-EM [Neal and Hinton, 1998, Cappé and Moulines, 2009];
- online dictionary learning [Mairal et al., 2010a].

Stochastic Majorization Minimization: SMM

Algorithm 3 Stochastic Majorization-Minimization Scheme

Input: $\theta_0 \in \Theta$ (initial estimate); N (number of iterations); $(w_n)_{n\geq 1}$, weights in $(0,1]$;
initialize the approximate surrogate: $\bar{g}_0 : \theta \mapsto \frac{\rho}{2}\|\theta - \theta_0\|_2^2$;
for $n = 1, \ldots, N$ **do**
  draw a training point $\mathbf{x}_n$;
  choose a surrogate function g_n of $f_n : \theta \mapsto \ell(\mathbf{x}_n, \theta)$ near θ_{n-1};
  update the approximate surrogate: $\bar{g}_n = (1 - w_n)\bar{g}_{n-1} + w_n g_n$;
  update the current estimate:

$$\theta_n \in \arg\min_{\theta \in \Theta} \bar{g}_n(\theta);$$

end for
Output: θ_N (current estimate);

Stochastic Majorization Minimization: SMM

Update Rule for Proximal Gradient Surrogate

$$\theta_n \leftarrow \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} w_n^i \left[\nabla f_i(\theta_{i-1})^\top \theta + \frac{L}{2} \|\theta - \theta_{i-1}\|_2^2 + \psi(\theta) \right]. \quad \text{(SMM)}$$

Other schemes in the literature [Duchi and Singer, 2009]:

$$\theta_n \leftarrow \arg\min_{\theta \in \Theta} \nabla f_n(\theta_{n-1})^\top \theta + \frac{1}{2\eta_n} \|\theta - \theta_{n-1}\|_2^2 + \psi(\theta), \quad \text{(FOBOS)}$$

or regularized dual averaging (RDA) of Xiao [2010]:

$$\theta_n \leftarrow \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_{i-1})^\top \theta + \frac{1}{2\eta_n} \|\theta\|_2^2 + \psi(\theta). \quad \text{(RDA)}$$

Stochastic Majorization Minimization: SMM

Theoretical Guarantees - Non-Convex Problems

under a set of reasonable assumptions,

- $f(\theta_n)$ almost surely converges;
- the function $\bar{g}_n$ asymptotically behaves as a first-order surrogates;
- we almost surely have asymptotic stationary point conditions.

Theoretical Guarantees - Convex Problems

for proximal gradient surrogates, we obtain similar expected rates as SGD with averaging [see Nemirovski et al., 2009, Polyak and Juditsky, 1992]: $O(1/n)$ for strongly convex problems, and $O(1/\sqrt{n})$ for convex ones.

Experimental Conclusions for ℓ_2-logistic Regression

Datasets

name	m	p	storage	size (GB)
alpha	250 000	500	dense	1
rcv1	781 265	47 152	sparse	0.95
covtype	581 012	54	dense	0.11
ocr	2 500 000	1 155	dense	23.1

for ℓ_2-logistic Regression

- Incremental and stochastic schemes were significantly faster than batch ones;
- MISO with heuristics was competitive with the state of the art (SAG, SGD, Liblinear);
- after one pass over the data, SMM quickly achieves a **low-precision** solution. For higher precision, MISO is prefered.
- **problems tested were large but relatively well conditioned**.

Stochastic DC programming

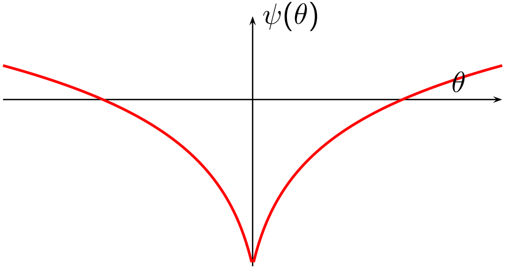
Consider a binary classification problem with enormous training data $(y_n, \mathbf{x}_n)$, with y_n in $\{-1, +1\}$ and $\mathbf{x}_n$ in $\mathbb{R}^p$. Assume that there exists a sparse linear model $y \approx \text{sign}(\theta^\top \mathbf{x}_i)$, learned by minimizing

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{(y,\mathbf{x})}[\log(1 + e^{-y\theta^\top \mathbf{x}})] + \lambda\psi(\theta).$$

Traditional choices for ψ: $\psi(\theta) = \|\theta\|_2^2$ or $\|\theta\|_1$.

Non-convex sparsity inducing penalty:

- $\psi(\theta) = \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon)$.



Stochastic DC programming

- upper-bound $f_n : \theta \mapsto \log(1 + e^{-y_n \theta^\top \mathbf{x}_n})$ by

$$\theta \mapsto f_n(\theta_{n-1}) + \nabla f_n(\theta_{n-1})^\top (\theta - \theta_{n-1}) + \frac{L}{2}\|\theta - \theta_{n-1}\|_2^2;$$

- upper-bound $\lambda \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon)$ by

$$\theta \mapsto \lambda \sum_{j=1}^{p} \frac{|\theta[j]|}{|\theta_{n-1}[j]| + \varepsilon}.$$

this is a stochastic reweighted-ℓ_1 algorithm [Candès et al., 2008].

Stochastic DC programming

- upper-bound $f_n : \theta \mapsto \log(1 + e^{-y_n \theta^\top \mathbf{x}_n})$ by

$$\theta \mapsto f_n(\theta_{n-1}) + \nabla f_n(\theta_{n-1})^\top (\theta - \theta_{n-1}) + \frac{L}{2}\|\theta - \theta_{n-1}\|_2^2;$$

- upper-bound $\lambda \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon)$ by

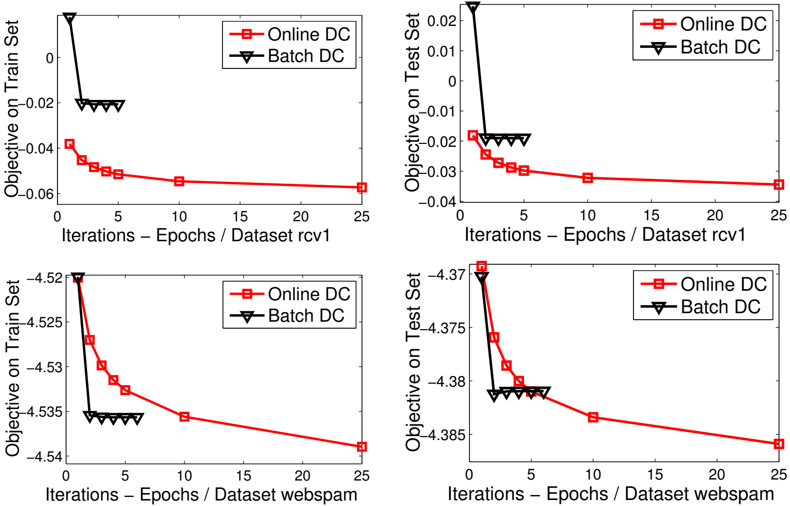
$$\theta \mapsto \lambda \sum_{j=1}^{p} \frac{|\theta[j]|}{|\theta_{n-1}[j]| + \varepsilon}.$$

this is a stochastic reweighted-ℓ_1 algorithm [Candès et al., 2008].

Datasets

name	N_{tr} (train)	N_{te} (test)	p	density (%)
rcv1	781 265	23 149	47 152	0.161
webspam	250 000	100 000	16 091 143	0.023

Stochastic DC programming



Online Structured Matrix Factorization

Consider some signals $\mathbf{x}$ in $\mathbb{R}^m$. We want to find a dictionary $\mathbf{D}$ in $\mathbb{R}^{m \times K}$. The quality of $\mathbf{D}$ is measured through the loss

$$\ell(\mathbf{x}, \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\alpha}\|_2^2.$$

Then, learning the dictionary amounts to solving

$$\min_{\mathbf{D} \in \mathcal{C}} \mathbb{E}_{\mathbf{x}} \left[\ell(\mathbf{x}, \mathbf{D}) \right] + \varphi(\mathbf{D}),$$

and we can use the proximal gradient surrogate.

Why is it a matrix factorization problem?

$$\min_{\mathbf{D} \in \mathcal{C}, \mathbf{A} \in \mathbb{R}^{K \times n}} \frac{1}{2n} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \sum_{i=1}^{n} \lambda_1 \|\boldsymbol{\alpha}_i\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\alpha}_i\|_2^2 + \varphi(\mathbf{D}).$$

Online Structured Matrix Factorization

- when $\mathcal{C} = \{\mathbf{D} \in \mathbb{R}^{m\times K} \;\text{ s.t. }\; \|\mathbf{d}_j\|_2 \leq 1\}$ and $\varphi = 0$, the problem is called **sparse coding** or **dictionary learning** [Olshausen and Field, 1997, Elad and Aharon, 2006]. We can use the upper-bound

$$\ell(\mathbf{x}_n, \mathbf{D}) \leq \frac{1}{2}\|\mathbf{x}_n - \mathbf{D}\boldsymbol{\alpha}_n\|_2^2 + \lambda_1\|\boldsymbol{\alpha}_n\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}_n\|_2^2,$$

  where

$$\boldsymbol{\alpha}_n \triangleq \operatorname*{arg\,min}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}_n - \mathbf{D}_{n-1}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2,$$

  and we obtain the online dictionary learning of Mairal et al. [2010a].

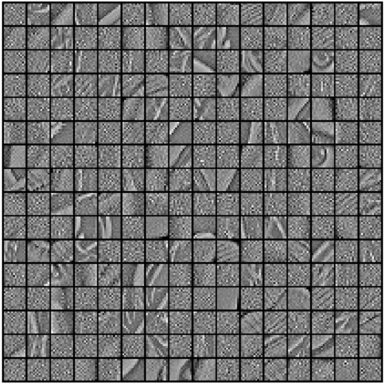
- non-negativity constraints can be easily added. It yields an online **nonnegative matrix factorization** algorithm.

- φ can be a function encouraging a particular structure in $\mathbf{D}$ [Jenatton et al., 2009].

Online Structured Matrix Factorization

Dictionary Learning on Natural Image Patches

Consider $n = 250\,000$ whitened natural image patches of size $m = 12 \times 12$. We learn a dictionary with $K = 256$ elements.

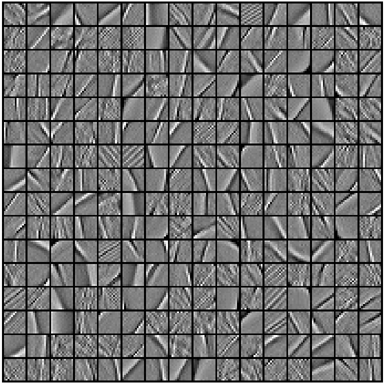


0s on an old laptop 1.2GHz dual-core CPU. (initialization)

Online Structured Matrix Factorization

Dictionary Learning on Natural Image Patches

Consider $n = 250\,000$ whitened natural image patches of size $m = 12 \times 12$. We learn a dictionary with $K = 256$ elements.

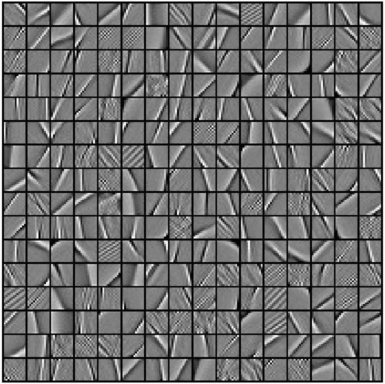


1.15s on an old laptop 1.2GHz dual-core CPU (0.1 pass)

Online Structured Matrix Factorization

Dictionary Learning on Natural Image Patches

Consider $n = 250\,000$ whitened natural image patches of size $m = 12 \times 12$. We learn a dictionary with $K = 256$ elements.

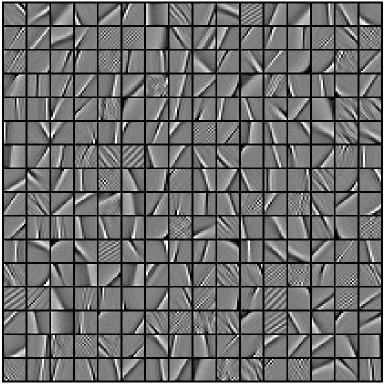


5.97s on an old laptop 1.2GHz dual-core CPU (0.5 pass)

Online Structured Matrix Factorization

Dictionary Learning on Natural Image Patches

Consider $n = 250\,000$ whitened natural image patches of size $m = 12 \times 12$. We learn a dictionary with $K = 256$ elements.

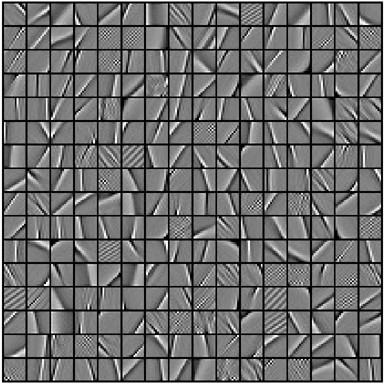


12.44s on an old laptop 1.2GHz dual-core CPU (1 pass)

Online Structured Matrix Factorization

Dictionary Learning on Natural Image Patches

Consider $n = 250\,000$ whitened natural image patches of size $m = 12 \times 12$. We learn a dictionary with $K = 256$ elements.

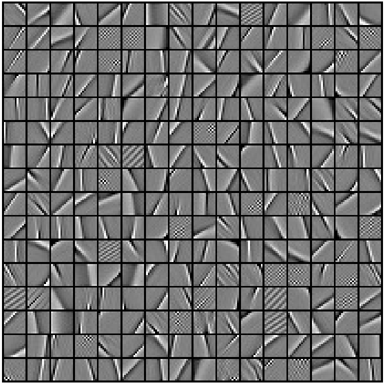


23.22s on an old laptop 1.2GHz dual-core CPU (2 passes)

Online Structured Matrix Factorization

Dictionary Learning on Natural Image Patches

Consider $n = 250\,000$ whitened natural image patches of size $m = 12 \times 12$. We learn a dictionary with $K = 256$ elements.



60.60s on an old laptop 1.2GHz dual-core CPU (5 passes)

Online Structured Matrix Factorization

With a structured regularization function φ [Jenatton et al., 2009]

$\varphi(\mathbf{D}) \triangleq \gamma_1 \sum_{j=1}^{K} \sum_{g \in \mathcal{G}} \max_{k \in g} |\mathbf{d}_j[k]| + \gamma_2 \|\mathbf{D}\|_F^2.$
The proximal operator of φ can be computed by using network flow optimization [Mairal et al., 2010b].

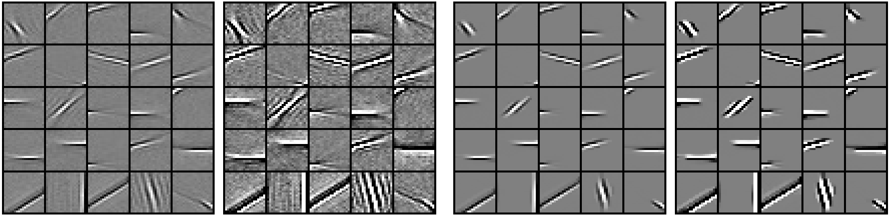


Figure: Left: subset of a larger dictionary obtained with ℓ_1; Right: subset obtained with φ after initialization with the dictionary on the left.

About 20 minutes per pass on the data on the 1.2GHz laptop CPU.

Conclusion

What we have done

- we have given a unified view of a large number of algorithms;
- ... and introduced new ones for large-scale optimization.

A take-home message

- our algorithms are likely to be useful for large-scale **non-convex** and possibly **non-smooth** problems.

Source Code

- code will be included in the toolbox SPAMS (C++ interfaced with Matlab, Python, R). `http://spams-devel.gforge.inria.fr/`;
- the online dictionary learning algorithm is already in SPAMS.

References I

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

D. Böhning and B. G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4): 641–663, 1988.

E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14: 877–905, 2008.

O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. 71(3):593–613, 2009.

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2010.

References II

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10: 2899–2934, 2009.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

T. Jebara and A. Choromanska. Majorization for CRFs and latent likelihoods. In *Advances in Neural Information Processing Systems*, 2012.

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, 2009. preprint arXiv:0904.3523v1.

Emtiyaz Khan, Ben Marlin, Guillaume Bouchard, and Kevin Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, 2010.

References III

K. Lange, D.R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. 9(1):1–20, 2000.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 2010a.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010b.

K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research*, (13):3441–3473, 2012.

R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, 89, 1998.

References IV

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. 19(4): 1574–1609, 2009.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, CORE, 2007.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37: 3311–3325, 1997.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

S. Shalev-Schwartz and T. Zhang. Proximal stochastic dual coordinate ascent. *preprint arXiv 1211.2717v1*, 2012.

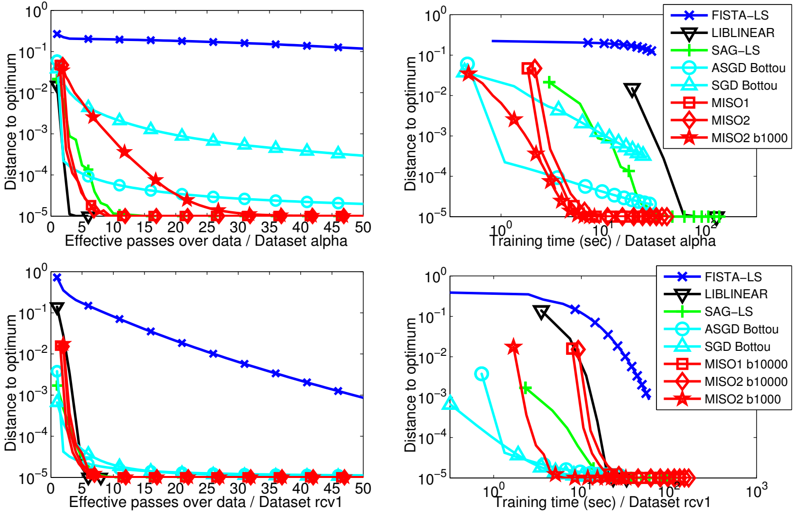
References V

S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 2008.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11: 2543–2596, 2010.

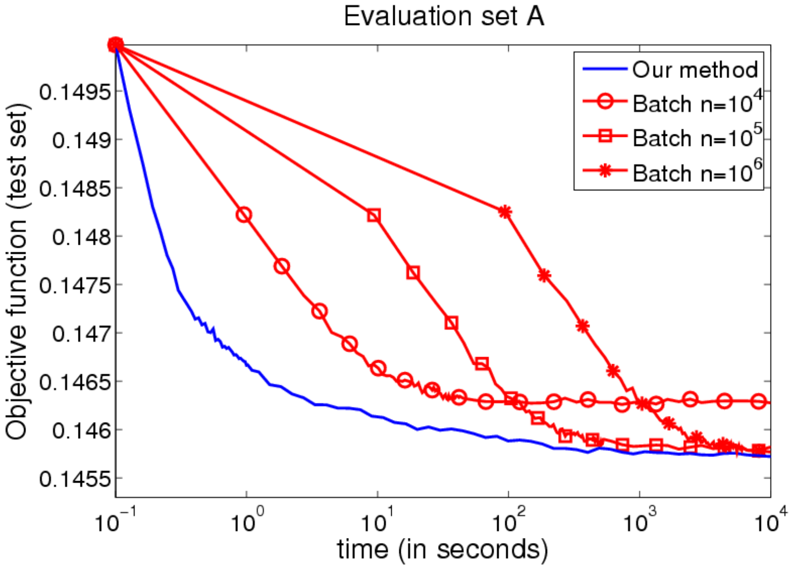
Performance of MISO for logistic-ℓ_2 regression

With preliminary version of SAG



Online Dictionary Learning

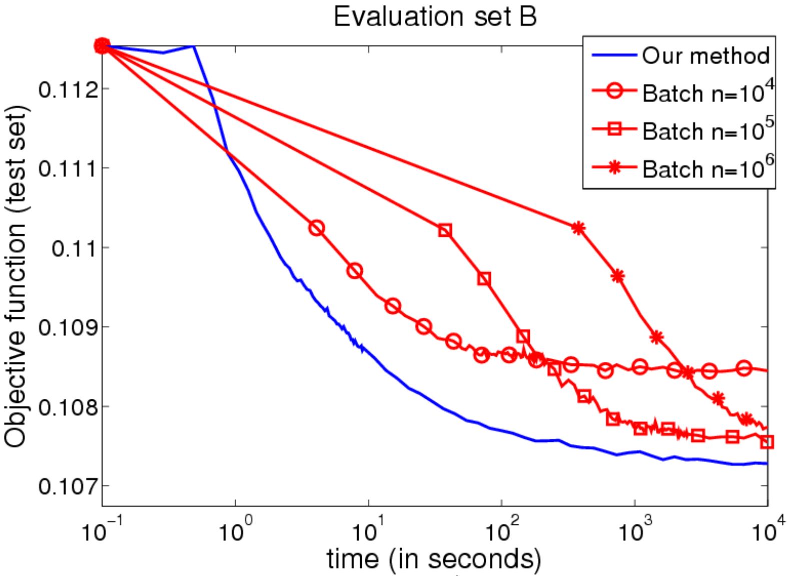
Experimental results, batch vs online



$m = 8 \times 8, \ k = 256$

Online Dictionary Learning

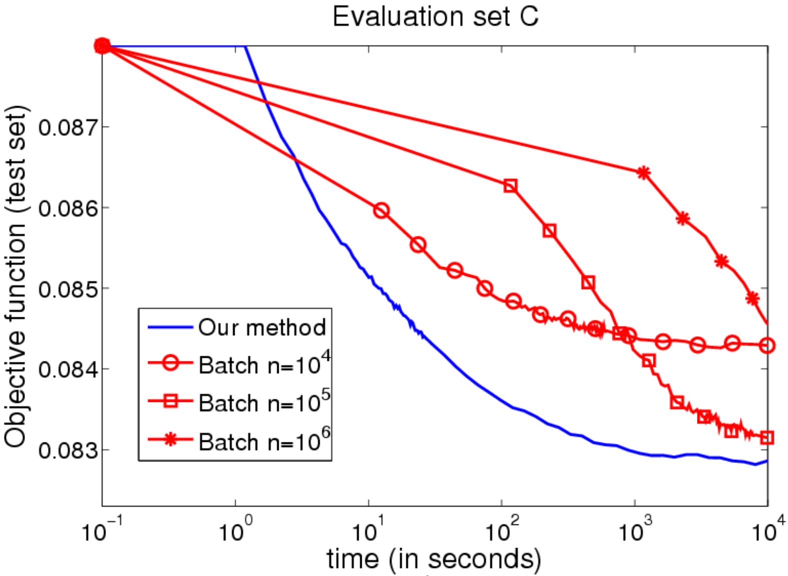
Experimental results, batch vs online



$m = 12 \times 12 \times 3,\ k = 512$

Online Dictionary Learning

Experimental results, batch vs online



$m = 16 \times 16$, $k = 1024$