Proximal Minimization by Incremental Surrogate Optimization (MISO)

(and a few variants)

Julien Mairal

Inria, Grenoble

ICCOPT, Tokyo, 2016

Motivation: large-scale machine learning

Minimizing large finite sums of functions

Given data points $\mathbf{x}_i$, $i = 1, \ldots, n$, learn some **model parameters** θ in $\mathbb{R}^p$ by minimizing

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta),$$

where ℓ measures the **data fit**, and ψ is a **regularization function**.

Minimizing expectations

If the amount of data is infinite, we may also need to minimize the **expected cost**

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

leading to a **stochastic optimization problem**.

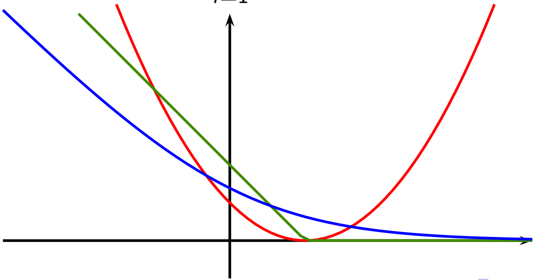
Motivation: large-scale machine learning

A few examples from the convex world

Ridge regression: $$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \langle \theta, \mathbf{x}_i\rangle)^2 + \frac{\lambda}{2}\|\theta\|_2^2.$$

Linear SVM: $$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\langle \theta, \mathbf{x}_i\rangle) + \frac{\lambda}{2}\|\theta\|_2^2.$$

Logistic regression: $$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i\langle \theta, \mathbf{x}_i\rangle}\right) + \frac{\lambda}{2}\|\theta\|_2^2.$$



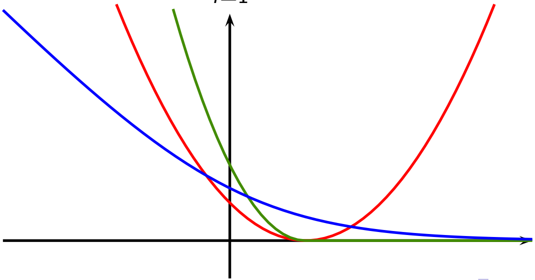
Motivation: large-scale machine learning

A few examples from the convex world

Ridge regression: $$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \langle \theta, \mathbf{x}_i \rangle)^2 + \frac{\lambda}{2}\|\theta\|_2^2.$$

Linear SVM: $$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\langle \theta, \mathbf{x}_i \rangle)^2 + \frac{\lambda}{2}\|\theta\|_2^2.$$

Logistic regression: $$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i\langle \theta, \mathbf{x}_i \rangle}\right) + \frac{\lambda}{2}\|\theta\|_2^2.$$



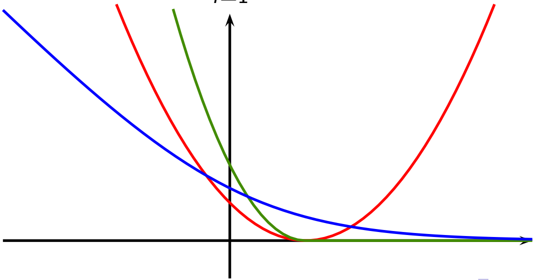
Motivation: large-scale machine learning

A few examples from the convex world

Ridge regression:
$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \langle \theta, \mathbf{x}_i \rangle)^2 + \lambda \|\theta\|_1.$$

Linear SVM:
$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\langle \theta, \mathbf{x}_i \rangle)^2 + \lambda \|\theta\|_1.$$

Logistic regression:
$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i\langle \theta, \mathbf{x}_i \rangle}\right) + \lambda \|\theta\|_1.$$



Methodology

We will consider optimization methods that iteratively build a **model** of the objective before updating the variable:

$$\theta_t \in \arg\min_{\theta \in \mathbb{R}^p} g_t(\theta),$$

where g_t is **easy to minimize** and exploits the objective structure: **large finite sum, expectation, (strong) convexity, composite**?

There is a large body of related work

- Kelley's and bundle methods;
- incremental and online EM algorithms;
- incremental and stochastic proximal gradient methods;
- variance-reduction techniques for minimizing finite sums.

[Neal and Hinton, 1998, Duchi and Singer, 2009, Bertsekas, 2011, Schmidt et al., 2013, Defazio et al., 2014a, Shalev-Shwartz and Zhang, 2012, Lan, 2012, 2015]...

Outline of the talk

1) stochastic majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

where ℓ is **not necessarily smooth or convex**.

2) incremental majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta).$$

$\Rightarrow$ The MISO algorithm for non-convex functions.

3) faster schemes for composite strongly-convex functions

$\Rightarrow$ Another MISO algorithm for strongly-convex functions.

4) ??

Outline of the talk

1) stochastic majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

where ℓ is **not necessarily smooth or convex**.

2) incremental majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta).$$

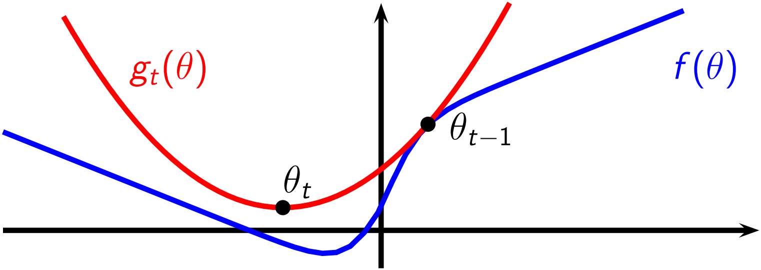
$\Rightarrow$ The MISO algorithm for non-convex functions.

3) faster schemes for composite strongly-convex functions

$\Rightarrow$ Another MISO algorithm for strongly-convex functions.

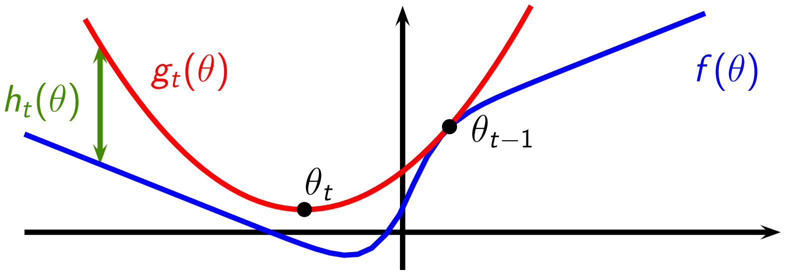
4) ??

Majorization-minimization principle



- Iteratively minimize locally tight upper bounds of the objective.
- The objective monotonically decreases.
- Under some assumptions, we get similar convergence rates as classical first-order approaches in the convex case.

Setting: first-order surrogate functions



- $g_t(\theta_t) \geq f(\theta_t)$ for θ_t in $\arg\min_{\theta \in \Theta} g_t(\theta)$;
- the **approximation error** $h_t \triangleq g_t - f$ is differentiable, and ∇h_t is L-Lipschitz. Moreover, $h_t(\theta_{t-1}) = 0$ and $\nabla h_t(\theta_{t-1}) = 0$;
- we may also need g_t to be strongly convex.

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**:
  $f = f' + \psi$ with f' smooth.

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

  Minimizing g amounts to one step of the forward-backward, ISTA, or proximal gradient descent algorithm.

[Nesterov, 2004, 2013, Beck and Teboulle, 2009, Wright et al., 2009]...

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 $\nabla^2 f$:

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

- Upper-bounds based on **Jensen's inequality**...

Theoretical guarantees of the basic MM algorithm

When using first-order surrogates,

- for **convex** problems: $f(\theta_t) - f^\star = O(L/t)$.
- for μ-**strongly convex** ones: $O((1-\mu/L)^t)$.
- for **non-convex** problems: $f(\theta_t)$ monotonically decreases and

$$\liminf_{t \to +\infty} \inf_{\theta \in \Theta} \frac{\nabla f(\theta_t, \theta - \theta_t)}{\|\theta - \theta_t\|_2} \geq 0, \tag{1}$$

  which we call asymptotic stationary point condition.

Directional derivative

$$\nabla f(\theta, \kappa) = \lim_{\varepsilon \to 0^+} \frac{f(\theta + \varepsilon \kappa) - f(\theta)}{\varepsilon}.$$

- when $\Theta = \mathbb{R}^p$ and f is smooth, (1) is equivalent to $\nabla f(\theta_t) \to 0$.

Stochastic majorization minimization [Mairal, 2013]

Assume that f is an expectation:

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

Recipe

- Draw **a single function** $f_t : \theta \mapsto \ell(\theta, \mathbf{x}_t)$ at iteration t;
- Choose a **first-order surrogate function** $\tilde{g}_t$ for f_t at θ_{t-1};
- **Update the model** $g_t = (1-w_t)g_{t-1} + w_t\tilde{g}_t$ with appropriate w_t;
- Update θ_t by minimizing g_t.

Related Work

- online-EM;
- online matrix factorization.

[Neal and Hinton, 1998, Mairal et al., 2010, Razaviyayn et al., 2013]...

Stochastic majorization minimization [Mairal, 2013]

Theoretical Guarantees - Non-Convex Problems

under a set of reasonable assumptions,

- $f(\theta_t)$ **almost surely converges**;
- the function g_t asymptotically behaves as a first-order surrogate;
- **asymptotic stationary point conditions** hold almost surely.

Theoretical Guarantees - Convex Problems

under a few assumptions, for proximal gradient surrogates, we obtain similar expected rates as SGD with averaging: $O(1/t)$ for **strongly convex problems**, $O(\log(t)/\sqrt{t})$ for **convex ones**.

The most interesting feature of this principle is probably the ability to deal with some non-smooth non-convex problems.

Stochastic majorization minimization [Mairal, 2013]

Update Rule for Proximal Gradient Surrogate

$$\theta_t \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta} \sum_{i=1}^{t} w_t^i \left[\nabla f_i(\theta_{i-1})^\top \theta + \frac{L}{2} \|\theta - \theta_{i-1}\|_2^2 + \psi(\theta) \right]. \tag{SMM}$$

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

$$\theta_t \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta} \nabla f_t(\theta_{t-1})^\top \theta + \frac{1}{2\eta_t} \|\theta - \theta_{t-1}\|_2^2 + \psi(\theta), \tag{FOBOS}$$

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

$$\theta_t \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta} \frac{1}{t} \sum_{i=1}^{t} \nabla f_i(\theta_{i-1})^\top \theta + \frac{1}{2\eta_t} \|\theta\|_2^2 + \psi(\theta). \tag{RDA}$$

or others...

Outline of the talk

1) stochastic majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

where ℓ is **not necessarily smooth or convex**.

2) incremental majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta).$$

$\Rightarrow$ The MISO algorithm for non-convex functions.

3) faster schemes for composite strongly-convex functions

$\Rightarrow$ Another MISO algorithm for strongly-convex functions.

4) ??

MISO (MM) for non-convex optimization [Mairal, 2015]

Assume that f splits into many components:

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f^i(\theta).$$

Recipe

- Draw at random **a single index** i_t at iteration t;
- Compute a **first-order surrogate** $g_t^{i_t}$ of f^{i_t} at θ_{t-1};
- **Incrementally update** the approximate surrogate

$$g_t \triangleq \frac{1}{n}\sum_{i=1}^{n} g_t^i = g_{t-1} + \frac{1}{n}(g_t^{i_t} - g_{t-1}^{i_t}).$$

- Update θ_t by minimizing g_t.

MISO (MM) for non-convex optimization [Mairal, 2015]

Theoretical Guarantees - Non-Convex Problems

same as the basic MM algorithm with probability one.

Theoretical Guarantees - Convex Problems

when using proximal gradient surrogates,

- for **convex problems**, $f(\hat{\theta}_t) - f^\star = O(nL/t)$.
- for μ-**strongly convex problems**, $f(\theta_t) - f^\star = O((1 - \mu/(nL))^t)$.

The computational complexity is the same as ISTA.

Related work for non-convex problems

- incremental EM;
- more specific incremental MM algorithms.

[Neal and Hinton, 1998, Ahn et al., 2006].

Outline of the talk

1) stochastic majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

where ℓ is **not necessarily smooth or convex**.

2) incremental majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta).$$

$\Rightarrow$ The MISO algorithm for non-convex functions.

3) faster schemes for composite strongly-convex functions

$\Rightarrow$ Another MISO algorithm for strongly-convex functions.

4) ??

MISO for μ-strongly convex smooth functions

Strong convexity provides simple **quadratic surrogate functions**:

$$g_t^i : \theta \mapsto f^i(\theta_{t-1}) + \nabla f^i(\theta_{t-1})^\top(\theta - \theta_{t-1}) + \frac{\mu}{2}\|\theta - \theta_{t-1}\|_2^2. \qquad (2)$$

This time, the model of the objective is a **lower bound**.

Proposition: MISO with lower bounds [Mairal, 2015]

When the functions f_i are μ-strongly convex, L-smooth, and non-negative, MISO with the surrogates (2) guarantees that

$$\mathbb{E}[f(\theta_t) - f^\star] \leq \left(1 - \frac{1}{3n}\right)^t nf^\star,$$

under the **"big data" condition** $n \geq 2L/\mu$.

Remark

- When $n \leq 2L/\mu$, the algorithm may diverge.

MISO for μ-strongly convex composite functions [Lin, Mairal, and Harchaoui, 2015]

First goal: allow a composite term ψ

$$f(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} f^i(\theta) + \psi(\theta),$$

by simply using the **composite lower-bounds**

$$g_t^i : \theta \mapsto f^i(\theta_{t-1}) + \nabla f^i(\theta_{t-1})^\top(\theta - \theta_{t-1}) + \frac{\mu}{2}\|\theta - \theta_{t-1}\|_2^2 + \psi(\theta). \quad (\star)$$

Second goal: remove the condition $n \geq 2L/\mu$

$$g_t^i : \theta \mapsto (1-\delta)g_{t-1}^i(\theta) + \delta(\star), \tag{3}$$

with $\delta = \min\left(1, \frac{\mu n}{2(L-\mu)}\right)$ instead of $\delta = 1$ previously.

MISO for μ-strongly convex composite functions [Lin, Mairal, and Harchaoui, 2015]

Convergence of MISO-prox

When the functions f_i are μ-strongly convex, L-smooth, MISO-prox with the surrogates (3) guarantees that

$$\mathbb{E}[f(\theta_t)] - f^* \leq \frac{1}{\tau}(1-\tau)^{t+1}\left(f(\theta_0) - g_0(\theta_0)\right) \quad \text{with} \quad \tau \geq \min\left\{\frac{\mu}{4L}, \frac{1}{2n}\right\}.$$

Furthermore, we also have fast convergence of the certificate

$$\mathbb{E}[f(\theta_t) - g_t(\theta_t)] \leq \frac{1}{\tau}(1-\tau)^{t}\left(f^* - g_0(\theta_0)\right).$$

MISO for μ-strongly convex composite functions [Lin, Mairal, and Harchaoui, 2015]

Relation with SDCA [Shalev-Shwartz and Zhang, 2012].

- Variant "5" of SDCA is identical to MISO-Prox with $\delta = \frac{\mu n}{L+\mu n}$;
- The construction is **primal**. The proof of convergence and the algorithm do not use duality, whereas SDCA is a dual ascent technique;
- $g_t(\theta_t)$ is a lower-bound of $f^\star$; it plays the same role as the dual lower bound in SDCA, but is **easier to evaluate**.

Another viewpoint about SDCA without duality [Shalev-Shwartz, 2015].

MISO for μ-strongly convex composite functions

We may now compare the **expected** complexity, using the fact that incremental algorithms require to **compute a single** ∇f^i per iteration.

	$\mu > 0$
grad. desc., ISTA, **MISO-MM**	$O\left(n\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right)$
FISTA, acc. grad. desc.	$O\left(n\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)\right)$
SVRG, SAG, SAGA, SDCA, **MISO**μ, Finito	$O\left(\max\left(n,\frac{L}{\mu}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$

SVRG, SAG, SAGA, SDCA, MISO, Finito improve upon FISTA when

$$\max\left(n,\frac{L}{\mu}\right) \leq n\sqrt{\frac{L}{\mu}} \quad \Leftrightarrow \sqrt{\frac{L}{\mu}} \leq n,$$

[Schmidt et al., 2013, Xiao and Zhang, 2014, Defazio et al., 2014a,b, Shalev-Shwartz and Zhang, 2012, Zhang and Xiao, 2015]

Outline of the talk

1) stochastic majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta),$$

where ℓ is **not necessarily smooth or convex**.

2) incremental majorization-minimization

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, \theta) + \psi(\theta).$$

$\Rightarrow$ The MISO algorithm for non-convex functions.

3) faster schemes for composite strongly-convex functions

$\Rightarrow$ Another MISO algorithm for strongly-convex functions.

4) ??

Can we do better? [Lin, Mairal, and Harchaoui, 2015]

SVRG, SAG, SAGA, SDCA, MISO, Finito improve upon FISTA **but they are not "accelerated" in the sense of Nesterov.**

[Beck and Teboulle, 2009, Nesterov, 2013, Shalev-Shwartz and Zhang, 2014, Lan, 2015, Agarwal and Bottou, 2015, Allen-Zhu, 2016]

Can we do better? [Lin, Mairal, and Harchaoui, 2015]

SVRG, SAG, SAGA, SDCA, MISO, Finito improve upon FISTA **but they are not "accelerated" in the sense of Nesterov.**

How to improve the previous complexities?

1. read classical paper about accelerated gradient methods;

[Beck and Teboulle, 2009, Nesterov, 2013, Shalev-Shwartz and Zhang, 2014, Lan, 2015, Agarwal and Bottou, 2015, Allen-Zhu, 2016]

Can we do better? [Lin, Mairal, and Harchaoui, 2015]

SVRG, SAG, SAGA, SDCA, MISO, Finito improve upon FISTA **but they are not "accelerated" in the sense of Nesterov.**

How to improve the previous complexities?

1. read classical paper about accelerated gradient methods;
2. stay in the room and listen to G. Lan's talk;

[Beck and Teboulle, 2009, Nesterov, 2013, Shalev-Shwartz and Zhang, 2014, Lan, 2015, Agarwal and Bottou, 2015, Allen-Zhu, 2016]

Can we do better? [Lin, Mairal, and Harchaoui, 2015]

SVRG, SAG, SAGA, SDCA, MISO, Finito improve upon FISTA **but they are not "accelerated" in the sense of Nesterov.**

How to improve the previous complexities?

1. read classical paper about accelerated gradient methods;
2. stay in the room and listen to G. Lan's talk;
3. **Listen to Hongzhou's Lin talk tomorrow**.

Tuesday, 2:45pm, room 5A

[Beck and Teboulle, 2009, Nesterov, 2013, Shalev-Shwartz and Zhang, 2014, Lan, 2015, Agarwal and Bottou, 2015, Allen-Zhu, 2016]

Conclusion

- a large class of **majorization-minimization** algorithms for **non-convex, possibly non-smooth, optimization**;
- fast algorithms for minimizing **large sums of convex functions** (using lower bounds).
- **see Hongzhou Lin's talk on acceleration tomorrow**.

Related publications

- J. Mairal. Optimization with First-Order Surrogate Functions. *ICML*, 2013.
- J. Mairal. Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization. *NIPS*, 2013.
- J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 2015;
- H. Lin, J. Mairal, and Z. Harchaoui. A Universal Catalyst for First-Order Optimization. *NIPS*, 2015;

Online Sparse Matrix Factorization

Consider some signals $\mathbf{x}$ in $\mathbb{R}^m$. We want to find a dictionary $\mathbf{D}$ in $\mathbb{R}^{m \times p}$. 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

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

Why is it a matrix factorization problem?

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

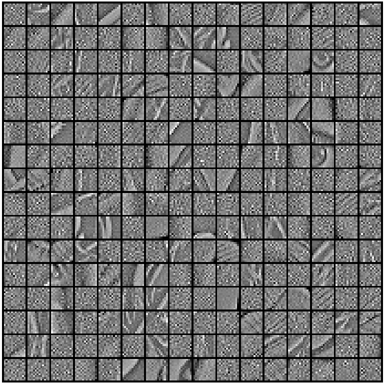
Online Sparse Matrix Factorization

- when $\mathcal{D} = \{\mathbf{D} \in \mathbb{R}^{m \times p} \ \text{ s.t. } \ \|\mathbf{d}_j\|_2 \leq 1\}$ and $\varphi = 0$, the problem is called **sparse coding** or **dictionary learning** [Olshausen and Field, 1996, Elad and Aharon, 2006, Mairal et al., 2010].
- 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., 2011].

Online Sparse 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.

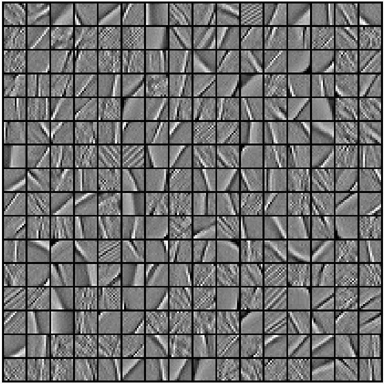


0s on an old laptop 1.2GHz dual-core CPU. (initialization)

Online Sparse 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.

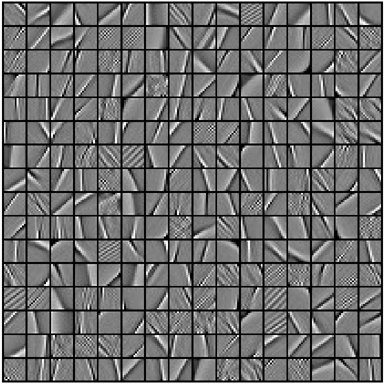


1.15s on an old laptop 1.2GHz dual-core CPU (0.1 pass)

Online Sparse 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.

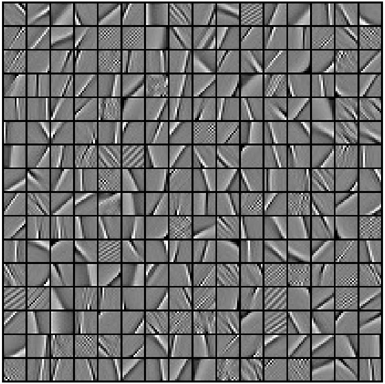


5.97s on an old laptop 1.2GHz dual-core CPU (0.5 pass)

Online Sparse 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.

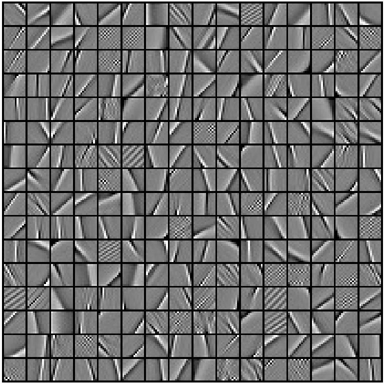


12.44s on an old laptop 1.2GHz dual-core CPU (1 pass)

Online Sparse 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.

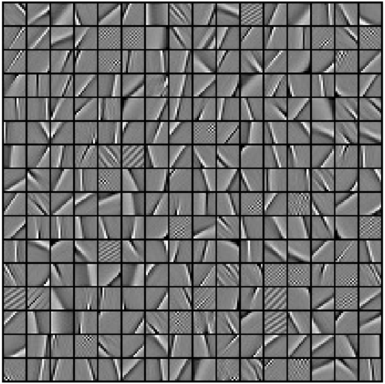


23.22s on an old laptop 1.2GHz dual-core CPU (2 passes)

Online Sparse 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)

References I

A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

Sangtae Ahn, Jeffrey A Fessler, Doron Blatt, and Alfred O Hero. Convergent incremental optimization transfer algorithms: Application to tomography. *IEEE Transactions on Medical Imaging*, 25(3):283–296, 2006.

Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *ArXiv 1603.05953*, 2016.

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. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010: 1–38, 2011.

References II

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014a.

A. J. Defazio, T. S. Caetano, and J. Domke. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014b.

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*, 15(12):3736–3745, 2006.

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12: 2777–2824, 2011.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

References III

Guanghui Lan. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, 2015.

J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2): 829–855, 2015.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11: 19–60, 2010.

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

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite objective function. *Mathematical Programming*, 140(1):125–161, 2013.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381: 607–609, 1996.

Meisam Razaviyayn, Maziar Sanjabi, and Zhi-Quan Luo. A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *arXiv preprint arXiv:1307.4457*, 2013.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

S. Shalev-Shwartz. Sdca without duality. *arxiv 1502:06177*, 2015.

S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. *arXiv:1211.2717*, 2012.

References V

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pages 1–41, 2014.

S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7): 2479–2493, 2009.

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

L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.