a Randomized Proximal Gradient Method with Structure-Adapted Sampling

Franck Iutzeler LJK, Univ. Grenoble Alpes

Journées SMAI MODE 2020
7-9 Sept. 2020

Structure	Regularization
sparsity	$r = \|\cdot\|_1$
anti-sparsity	$r = \|\cdot\|_\infty$
low rank	$r = \|\cdot\|_*$
⋮	⋮

Linear inverse problems: for a chosen regularization, we seek

$$x^\star \in \arg\min_x r(x) \qquad \text{such that } Ax = b$$

Regularized **Empirical Risk Minimization** problem:

$$\text{Find} \quad x^\star \in \arg\min_{x\in\mathbb{R}^n} \quad \underbrace{\mathcal{R}\left(x; \{a_i, b_i\}_{i=1}^m\right)}_{\text{obtained from statistical modeling}} + \underbrace{\lambda\ r(x)}_{\text{chosen regularization}}$$

$$\text{e.g. Lasso: Find} \quad x^\star \in \arg\min_{x\in\mathbb{R}^n} \quad \sum_{i=1}^m \tfrac{1}{2}(a_i^\top x - b_i)^2 + \lambda\ \|x\|_1$$

Regularization can improve statistical properties (generalization, stability, ...).

- ◇ Tibshirani: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (1996)
- ◇ Tibshirani *et al.*: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society (2004)
- ◇ Vaiter, Peyré, Fadili: Model consistency of partly smooth regularizers. IEEE Trans. on Information Theory (2017)

Composite minimization

$$\text{Find} \quad x^\star \in \arg\min_{x\in\mathbb{R}^n} \quad \mathcal{R}\left(x; \{a_i, b_i\}_{i=1}^m\right) \quad + \quad \lambda\, r(x)$$

$$\text{Find} \quad x^\star \in \arg\min_{x\in\mathbb{R}^n} \quad f(x) \quad + \quad g(x)$$

smooth — non-smooth

- f: differentiable surrogate of the empirical risk $\Rightarrow$ **Gradient**
  non-linear smooth function that depends on all the data
- g: non-smooth but chosen regularization $\Rightarrow$ **Proximity operator**
  non-differentiability on some manifolds implies structure on the solutions

$$\mathbf{prox}_{\gamma g}(u) = \arg\min_{y\in\mathbb{R}^n} \left\{ g(y) + \tfrac{1}{2\gamma}\|y-u\|_2^2 \right\}$$

closed form/easy for many regularizations:

- $g(x) = \|x\|_1$
- $g(x) = TV(x)$
- $g(x) = indicator_C(x)$

Natural optimization method: **proximal gradient**

$$\begin{cases} u_{k+1} = x_k - \gamma \nabla f(x_k) \\ x_{k+1} = \mathbf{prox}_{\gamma g}(u_{k+1}) \end{cases}$$

and its stochastic variants: proximal sgd, etc.

Example: LASSO

$$\begin{array}{rlccc} \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \mathcal{R}\left(x;\{a_i,b_i\}_{i=1}^m\right) & + & \lambda\ r(x) \\ \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \frac{1}{2}\left\|Ax-b\right\|_2^2 & + & \lambda\|x\|_1 \\ & & \text{smooth} & & \text{non-smooth} \end{array}$$

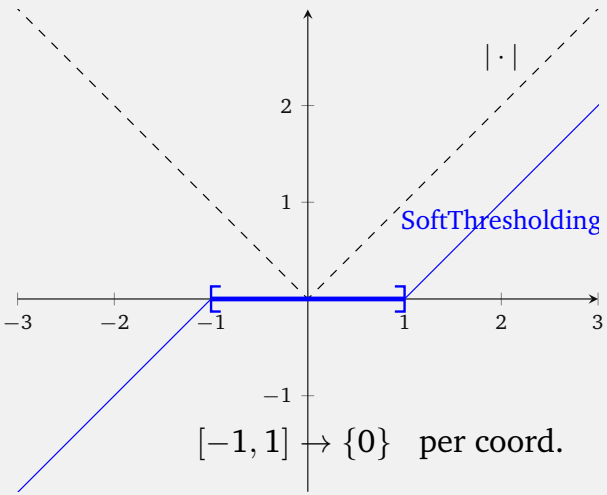
Coordinates	**Structure**	$\leftrightarrow$	**Optimality conditions**
$\forall i$	$x_i^\star = 0$	$\Leftrightarrow$	$A_i^\top (Ax^\star - b) \in [-\lambda, \lambda]$

Proximity Operator: per coordinate

$$\left[\mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u)\right]_i = \begin{cases} u_i - \lambda\gamma & \text{if } u_i > \lambda\gamma \\ 0 & \text{if } u_i \in [-\lambda\gamma;\lambda\gamma] \\ u_i + \lambda\gamma & \text{if } u_i < -\lambda\gamma \end{cases}$$

Proximal Gradient (aka ISTA):

$$\begin{cases} u_{k+1} = x_k - \gamma A^\top (Ax_k - b) \\ x_{k+1} = \mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u_{k+1}) \end{cases}$$



Example: LASSO

$$\begin{array}{rlccc} \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \mathcal{R}\left(x; \{a_i, b_i\}_{i=1}^m\right) & + & \lambda\ r(x) \\ \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \frac{1}{2}\left\|Ax-b\right\|_2^2 & + & \lambda\|x\|_1 \\ & & \text{smooth} & & \text{non-smooth} \end{array}$$

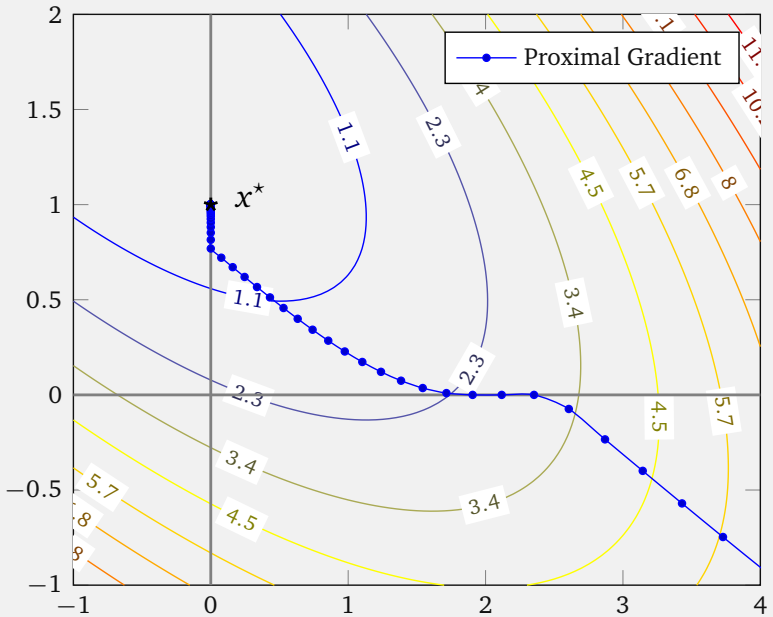
Coordinates	**Structure**	$\leftrightarrow$	**Optimality conditions**	$\leftrightarrow$	**Proximity operation**
$\forall i$	$x_i^\star = 0$	$\Leftrightarrow$	$A_i^\top (Ax^\star - b) \in [-\lambda, \lambda]$	$\Leftrightarrow$	$\left[\mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u^\star)\right]_i = 0$

$u^\star = x^\star - \gamma A^\top (Ax^\star - b)$

$$\left[\mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u)\right]_i = \begin{cases} u_i - \lambda\gamma & \text{if } u_i > \lambda\gamma \\ 0 & \text{if } u_i \in [-\lambda\gamma; \lambda\gamma] \\ u_i + \lambda\gamma & \text{if } u_i < -\lambda\gamma \end{cases}$$

Proximal Gradient (aka ISTA):

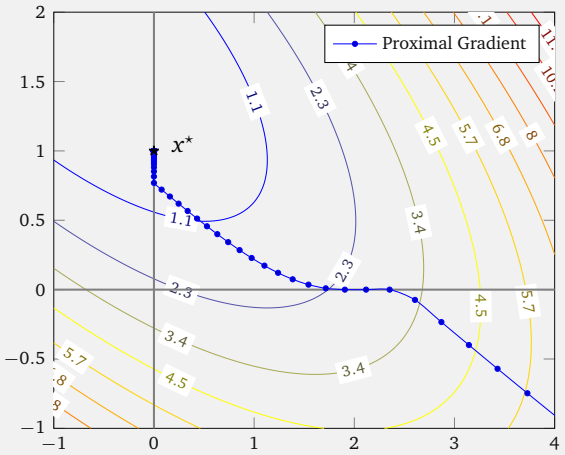
$$\begin{cases} u_{k+1} = x_k - \gamma A^\top (Ax_k - b) \\ x_{k+1} = \mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u_{k+1}) \end{cases}$$



Iterates (x_k) reach the **same structure as $x^\star$** in **finite** time!

Proximal Algorithms:

$$\begin{cases} u_{k+1} = x_k - \gamma \nabla f(x_k) \\ x_{k+1} = \mathbf{prox}_{\gamma g}(u_{k+1}) \end{cases}$$



g is **partly smooth** at x relative to a manifold $\mathcal{M} \ni x$ if:
- g is sharp across $\mathcal{M}$
- g is smooth along $\mathcal{M}$

> **project on manifolds**

Let $\mathcal{M}$ be a manifold and u_k such that

$$x_k = \mathbf{prox}_{\gamma g}(u_k) \in \mathcal{M} \quad \text{and} \quad \frac{u_k - x_k}{\gamma} \in \operatorname{ri} \partial g(x_k)$$

If g is partly smooth at x_k relative to $\mathcal{M}$, then

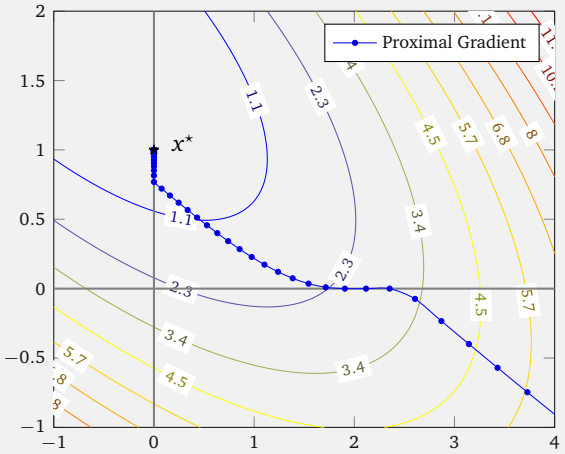
$$\mathbf{prox}_{\gamma g}(u) \in \mathcal{M}$$

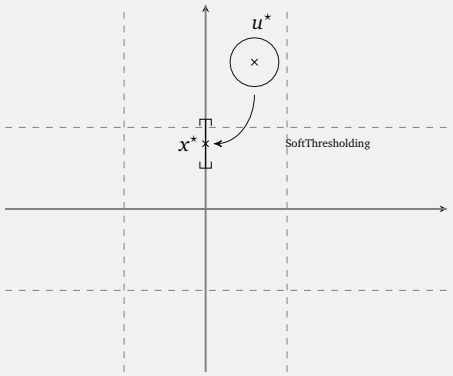
for any u close to u_k.

◇ Hare, Lewis: Identifying active constraints via partial smoothness and prox-regularity. Journal of Convex Analysis (2004)
◇ Daniilidis, Hare, Malick: Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. Optimization (2006)

Proximal Algorithms:

$$\begin{cases} u_{k+1} = x_k - \gamma \nabla f(x_k) \\ x_{k+1} = \mathbf{prox}_{\gamma g}(u_{k+1}) \end{cases}$$





> **project on manifolds**

> **identify the optimal structure**

Let (x_k) and (u_k) be a pair of sequences such that

$$x_k = \mathbf{prox}_{\gamma g}(u^k) \to x^\star = \mathbf{prox}_{\gamma g}(u^\star)$$

and $\mathcal{M}$ be a manifold. If $x^\star \in \mathcal{M}$ and

$$\exists \varepsilon > 0 \text{ such that for all } u \in \mathcal{B}(u^\star, \varepsilon),\ \mathbf{prox}_{\gamma g}(u) \in \mathcal{M} \qquad \text{(QC)}$$

holds, then, **after some finite but unknown time,** $x_k \in \mathcal{M}$.

◇ Lewis: Active sets, nonsmoothness, and sensitivity. SIAM Journal on Optimization (2002)

◇ Fadili, Malick, Peyré: Sensitivity analysis for mirror-stratifiable convex functions. SIAM Journal on Optimization (2018)

◇ Bareilles, I.: On the Interplay between Acceleration and Identification for the Proximal Gradient algorithm. Computational Optimization and Applications (2020)

> **Nonsmoothness** is actively studied in Numerical Optimization...

Subgradients, Partial Smoothness/prox-regularity, Bregman metrics, Error Bounds/Kurdyka-Łojasiewicz, etc.

- ◇ Hare, Lewis: Identifying active constraints via partial smoothness and prox-regularity. Journal of Convex Analysis (2004)
- ◇ Lemarechal, Oustry, Sagastizabal: The U-Lagrangian of a convex function. Transactions of the AMS (2000)
- ◇ Bolte, Daniilidis, Lewis: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization (2007)
- ◇ Chen, Teboulle: A proximal-based decomposition method for convex minimization problems. Mathematical Programming (1994)

> **Nonsmoothness** is actively studied in Numerical Optimization...
>
> Subgradients, Partial Smoothness/prox-regularity, Bregman metrics, Error Bounds/Kurdyka-Łojasiewicz, etc.

> ...but **often suffered** because of lack of structure/expression.
>
> Bundle methods, Gradient Sampling, Smoothing, Inexact proximal methods, etc.

- ◇ Nesterov: Smooth minimization of non-smooth functions. Mathematical Programming (2005)
- ◇ Burke, Lewis, Overton: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM Journal on Optimization (2005)
- ◇ Solodov, Svaiter: A hybrid projection-proximal point algorithm. Journal of convex analysis (1999)
- ◇ de Oliveira, Sagastizábal: Bundle methods in the XXIst century: A bird’s-eye view. Pesquisa Operacional (2014)

> **Nonsmoothness** is actively studied in Numerical Optimization...
Subgradients, Partial Smoothness/prox-regularity, Bregman metrics, Error Bounds/Kurdyka-Łojasiewicz, etc.

> ...but **often suffered** because of lack of structure/expression.
Bundle methods, Gradient Sampling, Smoothing, Inexact proximal methods, etc.

> For **Machine Learning objectives**, it can often be **harnessed**

- Explicit/“proximable” regularizations ℓ_1, nuclear norm
- We *know* the *expressions* and *activity* of sought structures sparsity, rank

◇ Bach, et al.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning (2012)
◇ Massias, Salmon, Gramfort: Celer: a fast solver for the lasso with dual extrapolation. ICML (2018)
◇ Liang, Fadili, Peyré: Local linear convergence of forward–backward under partial smoothness. NeurIPS (2014)
◇ O’Donoghue, Candes: Adaptive restart for accelerated gradient schemes. Foundations of computational mathematic (2015)

$$\begin{array}{rlrcl} \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \mathcal{R}\left(x;\{a_i,b_i\}_{i=1}^m\right) & + & \lambda\ r(x) \\ \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & f(x) & + & g(x) \\ & & \text{smooth} & & \text{non-smooth} \end{array}$$

A reason why the nonsmoothness of ML problems can be leveraged is that we *know* the *expressions* and *activity* of sought structures.

Mathematically, we can design a *lookout collection* $\mathsf{C} = \{\mathcal{M}_1, .., \mathcal{M}_p\}$ of closed sets such that:

(i) we have a projection mapping $\text{proj}_{\mathcal{M}_i}$ onto $\mathcal{M}_i$ for all i;

(ii) $\mathbf{prox}_{\gamma g}(u)$ is a singleton and can be computed explicitly for any u and γ;

(iii) upon computation of $x = \mathbf{prox}_{\gamma g}(u)$, we know if $x \in \mathcal{M}_i$ or not for all i.

$\Rightarrow$ **Structure can be directly observed**.

Example: Sparse structure and $g = \|\cdot\|_1, \|\cdot\|_{0.5}^{0.5}, \|\cdot\|_0, \ldots$

$$\mathsf{C} = \{\mathcal{M}_1, \ldots, \mathcal{M}_n\} \quad \text{with } \mathcal{M}_i = \{x \in \mathbb{R}^n : x_i = 0\}$$

> Identification of proximal algorithms

$\Rightarrow$ **After some finite time, the iterates reach the optimal structure**.

> Problems with noticeable strucure

$\Rightarrow$ **Structure can be directly observed**.

Let's use the structure progressively uncovered in proximal gradient!

not necessarily close to $x^\star$, no explicit or bad dual, when screening methods are difficult to implement

> Identification of proximal algorithms

⇒ **After some finite time, the iterates reach the optimal structure**.

> Problems with noticeable strucure

⇒ **Structure can be directly observed**.

Let's use the structure progressively uncovered in proximal gradient!

not necessarily close to $x^\star$, no explicit or bad dual, when screening methods are difficult to implement

Idea: Define $\mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_k \in \mathcal{M}_i} \mathcal{M}_i$ and $\mathcal{M}^\star := \mathbb{R}^n \bigcap_{i:x^\star \in \mathcal{M}_i} \mathcal{M}_i$, then:

$\mathcal{M}_k \subset \mathbb{R}^n$ partial identif/screening and $\mathcal{M}_k = \mathcal{M}^\star$ after some finite time identification

Can we reduce the dimension of the problem *on the fly* using $\mathcal{M}_k$?

Proximal Gradient methods with Adaptive Subspace Sampling

Dmitry Grishchenko F. I. Jerôme Malick

PhD student

Univ. Grenoble Alpes

CNRS and LJK

to appear in Mathematics of Operations Research

`https://arxiv.org/abs/2004.13356`

Disclaimer: We assume that the identified manifolds are linear subspaces eg: $\|Dx\|_1$ but not separability of g.

$$\begin{cases} y_k &= x_k - \gamma \nabla f(x_k) \\ z_k &= y_k \\ x_{k+1} &= \mathbf{prox}_{\gamma g}(z_k) \end{cases}$$

> Vanilla Proximal gradient identifies but does not use it
full gradient computed at each iteration

Example: Sparse structure and $g = \|\cdot\|_1$

$$\mathcal{C} = \{\mathcal{M}_1, \dots, \mathcal{M}_n\} \quad \text{with } \mathcal{M}_i = \{x \in \mathbb{R}^n : x_i = 0\}$$

Disclaimer: We assume that the identified manifolds are linear subspaces eg: $\|Dx\|_1$ but not separability of g.

$$\begin{cases} & \text{Observe } \mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_k \in \mathcal{M}_i} \quad \mathcal{M}_i \\ y_k &= \quad x_k - \gamma \nabla f(x_k) \\ z_k &= \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\ x_{k+1} &= \textbf{prox}_{\gamma g}(\quad z_k) \end{cases}$$

> Direct Use of Identification may not converge
>
> eg: starting with 0

Example: Sparse structure and $g = \|\cdot\|_1$

$$\mathcal{C} = \{\mathcal{M}_1, \dots, \mathcal{M}_n\} \quad \text{with } \mathcal{M}_i = \{x \in \mathbb{R}^n : x_i = 0\}$$
$$\mathcal{M}_k = \{x \in \mathbb{R}^n : x_i = 0 \text{ if } x_{i,k} = 0\}$$

Disclaimer: We assume that the identified manifolds are linear subspaces eg: $\|Dx\|_1$ but not separability of g.

$$\begin{cases} & \text{Observe } \mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_k \in \mathcal{M}_i} (\xi_{k,i}\mathcal{M}_i + (1-\xi_{k,i})\mathbb{R}^n) \text{ for } \xi_{k,i} \sim \mathcal{B}(p) \\ y_k &= x_k - \gamma \nabla f(x_k) \\ z_k &= \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\ x_{k+1} &= \mathbf{prox}_{\gamma g}(z_k) \end{cases}$$

> Mixing Identification and Randomized "coordinate" descent biases convergence issues

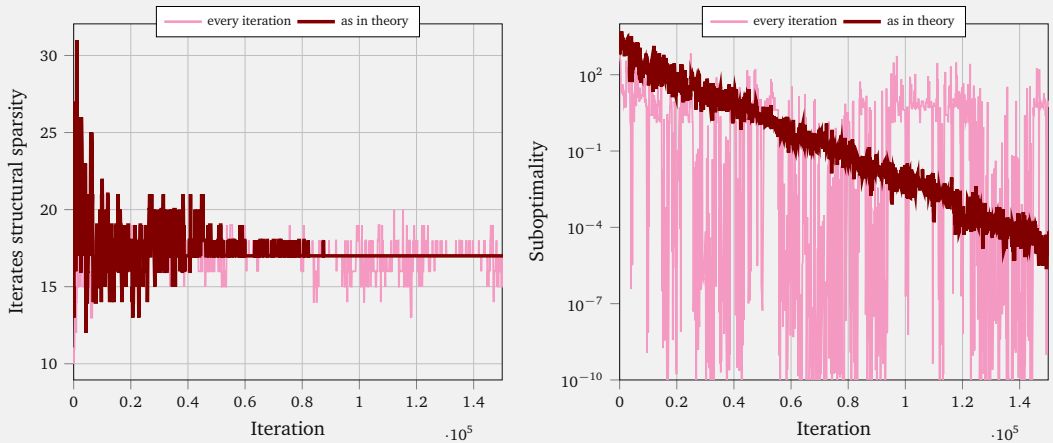
Example: Sparse structure and $g = \|\cdot\|_1$

$$\mathcal{C} = \{\mathcal{M}_1, \dots, \mathcal{M}_n\} \quad \text{with } \mathcal{M}_i = \{x \in \mathbb{R}^n : x_i = 0\}$$
$$\mathcal{M}_k = \{x \in \mathbb{R}^n : x_i = 0 \text{ if } x_{i,k} = 0 \quad \text{for some } i\}$$

Disclaimer: We assume that the identified manifolds are linear subspaces eg: $\|Dx\|_1$ but not separability of g.

$$
\begin{cases}
 & \text{Observe } \mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_k \in \mathcal{M}_i} (\xi_{k,i}\mathcal{M}_i + (1-\xi_{k,i})\mathbb{R}^n) \text{ for } \xi_{k,i} \sim \mathcal{B}(p) \\
 & \text{and compute } \mathbf{P}_k = \mathbb{E}\text{proj}_{\mathcal{M}_k} \text{ and } Q_k = (\mathbf{P}_k)^{-1/2} \\
y_k & = Q_k\left(x_k - \gamma \nabla f(x_k)\right) \\
z_k & = \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\
x_{k+1} & = \mathbf{prox}_{\gamma g}(Q_k^{-1} z_k)
\end{cases}
$$

> Unbiasing with $Q_k = (\mathbb{E}\text{proj}_{\mathcal{M}_k})^{-1/2}$ works *after identification*
but before... no, which prevents identification...

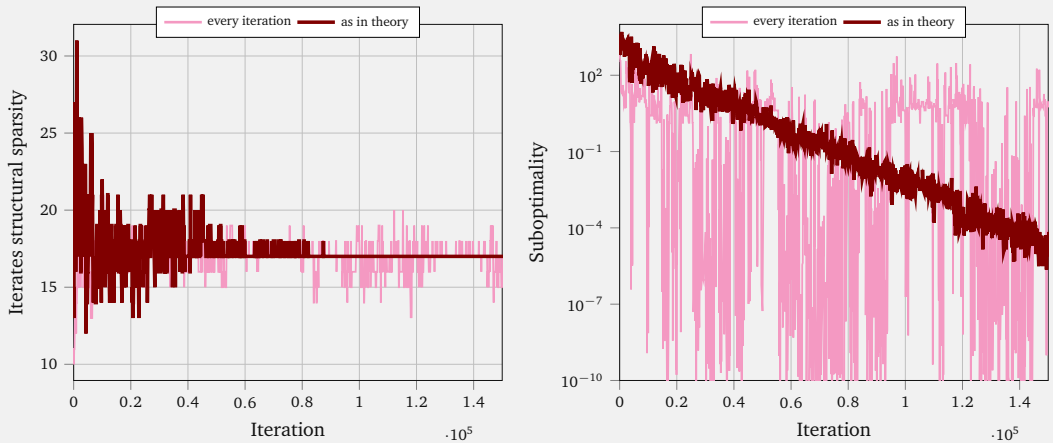


Disclaimer: We assume that the identified manifolds are linear subspaces eg: $\|Dx\|_1$ but not separability of g.

$$
\begin{cases}
 & \text{Observe } \mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_\ell \in \mathcal{M}_i} (\xi_{k,i}\mathcal{M}_i + (1-\xi_{k,i})\mathbb{R}^n) \text{ for } \xi_{k,i} \sim \mathcal{B}(p) \\
 & \text{and compute } \mathbf{P}_k = \mathbb{E}\text{proj}_{\mathcal{M}_k} \text{ and } Q_k = (\mathbf{P}_k)^{-1/2} \textit{sometimes}, \text{ else } \mathcal{M}_k = \mathcal{M}_{k-1} \\
y_k & = Q_k\left(x_k - \gamma \nabla f(x_k)\right) \\
z_k & = \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\
x_{k+1} & = \mathbf{prox}_{\gamma g}(Q_k^{-1} z_k)
\end{cases}
$$

> Structure adaptation can be performed at *some* iterations

depends on the *amount of change* $\|Q_{k-1}Q_k^{-1}\|$ and *harshness* of the sparsification $\lambda_{\min}(Q_k)$



$$\begin{cases} & \mathcal{M}_k = \mathbb{R}^n \bigcap_i (\xi_{k,i}\mathcal{M}_i + (1-\xi_{k,i})\mathbb{R}^n) \text{ for } \xi_{k,i} \sim \mathcal{B}(p)(iid) \\ & \text{Compute (once) } \mathbf{P} = \mathbb{E}\text{proj}_{\mathcal{M}_k} \text{ and } Q = (\mathbf{P})^{-1/2} \\ y_k & = Q\left(x_k - \gamma \nabla f(x_k)\right) \\ z_k & = \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\ x_{k+1} & = \mathbf{prox}_{\gamma g}(Q^{-1} z_k) \end{cases}$$

Theorem *Let f be L-smooth and μ-strongly convex and g be convex, lsc. Take any $\gamma \in (0, 2/(\mu+L)]$. Then, the sequence (x_k) converges almost surely to the minimizer $x^\star$ of $f+g$ and*

$$\mathbb{E}[\|x_k - x^\star\|^2] \leq \left(1 - \lambda_{\min}(\mathbf{P}) \frac{2\gamma\mu L}{\mu + L}\right)^k C$$

Proof. *Let $z^\star = y^\star = Q(x^\star - \gamma \nabla f(x^\star))$,*

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^\star\|^2 | \mathcal{F}_{k-1}] &\leq \mathbb{E}[\|z_k - z^\star\|^2_{\mathbf{P}} | \mathcal{F}_{k-1}] \\ &\leq \mathbb{E}[\|z_k - z^\star\|^2 | \mathcal{F}_{k-1}] \\ &\leq \|z_{k-1} - z^\star\|^2 + \mathbb{E}[\|y_k - y^\star\|^2_{\mathbf{P}} | \mathcal{F}_{k-1}] - \|z_{k-1} - z^\star\|^2_{\mathbf{P}} \\ &\leq \left(1 - \lambda_{\min}(\mathbf{P}) \frac{2\gamma\mu L}{\mu + L}\right) \|z_{k-1} - z^\star\|^2 \end{aligned}$$

$$\begin{cases} \text{if adaptation,} & \ell \leftarrow \ell + 1, x^{base} \leftarrow x_{k-c_\ell} \\ \text{(upd. space changes,} & \mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_i^{base}=0} (\xi_{k,i}\mathcal{M}_i + (1-\xi_{k,i})\mathbb{R}^n) \text{ for } \xi_{k,i} \sim \mathcal{B}(p)(iid) \\ \text{but same dist. whithin)} & \text{Compute (at each adapt.) } \mathbf{P}_\ell = \mathbb{E}\text{proj}_{\mathcal{M}_\ell} \text{ and } Q_\ell = (\mathbf{P}_\ell)^{-1/2} \\ y_k & = Q_\ell \left(x_k - \gamma \nabla f(x_k)\right) \\ z_k & = \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\ x_{k+1} & = \mathbf{prox}_{\gamma g}(Q_\ell^{-1} z_k) \end{cases}$$

Theorem *Let f be L-smooth and μ-strongly convex and g be convex, lsc. Take any $\gamma \in (0, 2/(\mu+L)]$. Consider the following adaptation strategy:*

1) If the structure of x_k changes, choose a new sampling $x_k \to x^{base}$, $\lambda_{\min}(\mathbf{P}_\ell) \geq \lambda$

2) Compute c_ℓ so that $\|Q_\ell Q_{\ell-1}^{-1}\|^2 \left(1 - \lambda_{\min}(\mathbf{P}_{\ell-1}) \frac{2\gamma\mu L}{\mu+L}\right)^{c_\ell} \leq \left(1 - \lambda \frac{\gamma\mu L}{\mu+L}\right)$

3) Apply the new sampling after c_ℓ iterations

Then, the sequence (x_k) converges almost surely to the minimizer $x^\star$ of $f+g$ and

$$\mathbb{E}[\|x_k - x^\star\|^2] \leq \left(1 - \lambda \frac{\gamma\mu L}{\mu+L}\right)^{\ell} C$$

where ℓ is the number of adaptations before k.

$$\begin{cases} \text{if adaptation,} & \ell \leftarrow \ell + 1, x^{base} \leftarrow x_{k-c_\ell} \\ \text{(upd. space changes,} & \mathcal{M}_k = \mathbb{R}^n \bigcap_{i:x_i^{base}=0} (\xi_{k,i}\mathcal{M}_i + (1-\xi_{k,i})\mathbb{R}^n) \text{ for } \xi_{k,i} \sim \mathcal{B}(p)(iid) \\ \text{but same dist. whithin)} & \text{Compute (at each adapt.) } \mathbf{P}_\ell = \mathbb{E}\text{proj}_{\mathcal{M}_\ell} \text{ and } Q_\ell = (\mathbf{P}_\ell)^{-1/2} \\ y_k & = Q_\ell \left(x_k - \gamma \nabla f(x_k)\right) \\ z_k & = \text{proj}_{\mathcal{M}_k}(y_k) + \text{proj}^{\perp}_{\mathcal{M}_k}(z_{k-1}) \\ x_{k+1} & = \mathbf{prox}_{\gamma g}(Q_\ell^{-1} z_k) \end{cases}$$

1) If the structure of x_k changes, choose a new sampling $x_k \to x^{base}$, $\lambda_{\min}(\mathbf{P}_\ell) \geq \lambda$
2) Compute c_ℓ so that $\|Q_\ell Q_{\ell-1}^{-1}\|^2 \left(1 - \lambda_{\min}(\mathbf{P}_{\ell-1}) \frac{2\gamma\mu L}{\mu+L}\right)^{c_\ell} \leq \left(1 - \lambda \frac{\gamma\mu L}{\mu+L}\right)$
3) Apply the new sampling after c_ℓ iterations

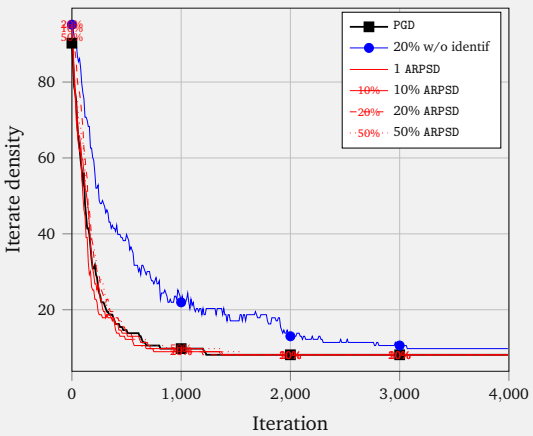
We “compensate” as along as the structure changes

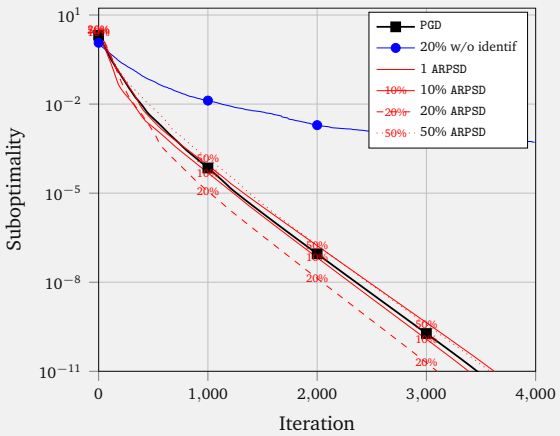
Theorem *Under the same assumptions as before + (QC), the rate improves to*

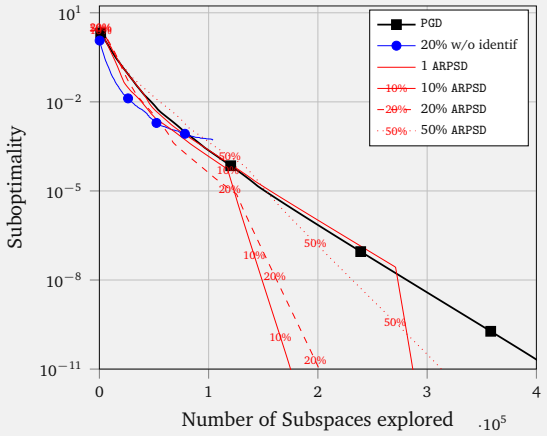
$$\mathbb{E}[\|x_k - x^\star\|^2] \leq \left(1 - \lambda_{\min}(\mathbf{P}^\star) \frac{2\gamma\mu L}{\mu + L}\right)^k C$$

linear in iterations (not adapt.) with a rate depending on the final structure.

TV-reg. logistic regression on a1a (1605×143), 90% final *jump* sparsity







- Iterate structure enforced by nonsmooth regularizers can be used to adapt the selection probabilities of coordinate descent/sketching;
- Before identification, adaptation *has to be moderate*;
- The cost of adaption may be big (in iterations and computation) but this can be mitigated in many practical cases.

- Machine Learning problems often have a *noticeable structure*;
- We can design a *lookout collection* $\mathsf{C} = \{\mathcal{M}_1, .., \mathcal{M}_p\}$ of sets: (i) with easy projections; (ii) identified by proximity operations; (iii) we *know* if these sets are identified or not;
- This structure can/should be harnessed but may be tricky before identification.

▷ Malick & I.: *Nonsmoothness in Machine Learning: specific structure,proximal identification, and applications*, review/pedagogical paper coming hopefully soon

Thanks to ANR JCJC *STROLL* & IDEX UGA IRS *DOLL* & PGMO

Thank you! – Franck IUTZELER `http://www.iutzeler.org`