Proximal Identification and Applications

Jérôme MALICK

CNRS, Lab. J. Kuntzmann, Grenoble (France)

Workshop Optimization for Machine Learning – Luminy – March 2020

talk based on materiel from joint work with

G. Peyré | J. Fadili | G. Garrigos | F. Iutzeler | D. Grishchenko



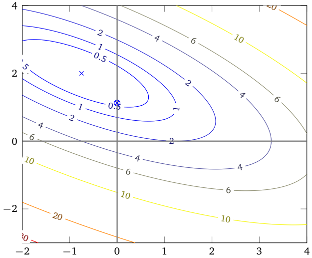


Example of stability

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2}\|Ax - y\|^2 \; + \; \lambda\|x\|_1 \qquad \text{(LASSO)}$$

Stability: the support of optimal solutions is stable under small perturbations

Illustration (on an instance with $d = 2$)

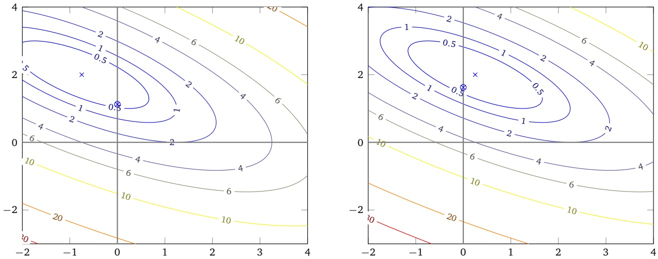


Example of stability

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2}\|Ax - y\|^2 \;+\; \lambda\|x\|_1 \qquad \text{(LASSO)}$$

Stability: the support of optimal solutions is stable under small perturbations

Illustration (on an instance with $d = 2$)

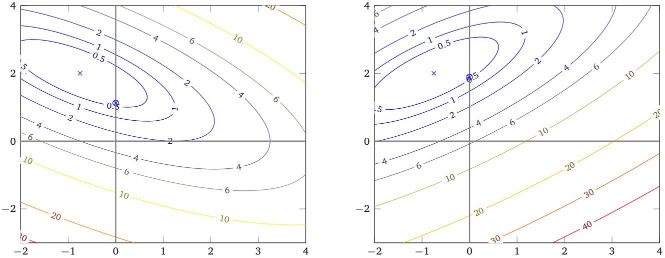


Example of stability

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2}\|Ax - y\|^2 \ + \ \lambda\|x\|_1 \qquad \text{(LASSO)}$$

Stability: the support of optimal solutions is stable under small perturbations

Illustration (on an instance with $d = 2$)

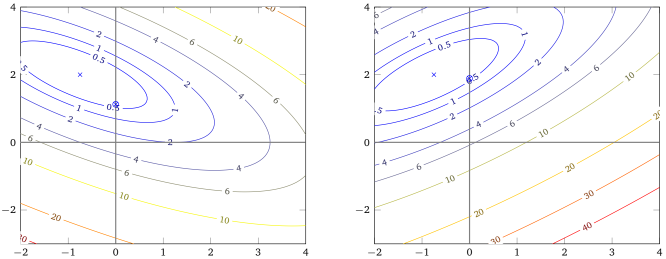


Example of stability

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2}\|Ax - y\|^2 \ + \ \lambda\|x\|_1 \qquad \text{(LASSO)}$$

Stability: the support of optimal solutions is stable under small perturbations

Illustration (on an instance with $d = 2$)



More generally: [Lewis '02] sensitivity analysis of partly-smooth functions

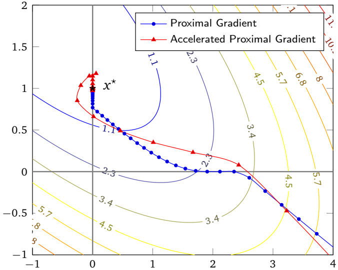
(remind Clarice's talk, this morning)

Example of identification

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2}\|Ax - y\|^2 \; + \; \lambda\|x\|_1 \qquad \text{(LASSO)}$$

Identification: (proximal-gradient) algorithms produce iterates...

...that eventually have the same support as the optimal solution



Runs of two proximal-gradient algos

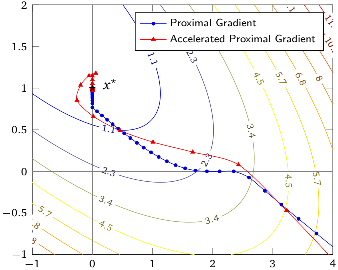
(same instance with $d = 2$)

Example of identification

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2}\|Ax - y\|^2 \; + \; \lambda\|x\|_1 \qquad \text{(LASSO)}$$

Identification: (proximal-gradient) algorithms produce iterates...

...that eventually have the same support as the optimal solution



Runs of two proximal-gradient algos

(same instance with $d = 2$)

Well-studied, see e.g. [Bertsekas '76], [Wright '96], [Lewis Drusvyatskiy '13]...

Outline

1. General stability of regularized problems
2. Enlarged identification of proximal algorithms
3. Application: communication-efficient federated learning
4. Application: model consistency for regularized least-squares

Outline

Stability or sensitivity analysis

Parameterized composite optimization problem (smooth + nonsmooth)

$$\min_{x\in\mathbb{R}^d} \quad F(x,p) + R(x),$$

Typically nonsmooth R traps solutions in low-dimensional manifolds

Stability: Optimal solutions lie on a manifold: $\quad x^\star(p) \in M \quad$ for $p \sim p_0$

Studied in e.g. [Hare Lewis '10] [Vaiter *et al* '15] [Liang *et al* '16]...

Example 1: $R = \|\cdot\|_1$, $\mathrm{supp}\big(x^\star(p)\big) = \mathrm{supp}\big(x^\star(p_0)\big)$

Stability or sensitivity analysis

Parameterized composite optimization problem (smooth + nonsmooth)

$$\min_{x \in \mathbb{R}^d} \quad F(x, p) + R(x),$$

Typically nonsmooth R traps solutions in low-dimensional manifolds

Stability: Optimal solutions lie on a manifold: $x^\star(p) \in M$ for $p \sim p_0$

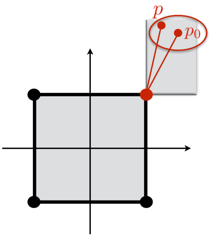
Studied in e.g. [Hare Lewis '10] [Vaiter *et al* '15] [Liang *et al* '16]...

Example 1: $R = \|\cdot\|_1$, $\mathrm{supp}\big(x^\star(p)\big) = \mathrm{supp}\big(x^\star(p_0)\big)$

Example 2: $R = \iota_{\mathbb{B}_\infty}$ (indicator function)

projection onto the ℓ_∞ ball

Many examples in machine learning...

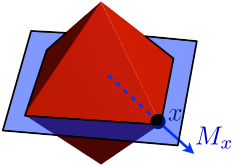


Structure of nonsmooth regularizers

Many of the regularizers used in machine learning or image processing have a strong primal-dual structure (“mirror-stratifiable” [Fadili, M., Peyré '18])
...that can be exploit to get (enlarged) stability/identification results

Examples: (associated unit ball and low-dimensional manifold where x belongs)

- $R = \|\cdot\|_1$ (and $\|\cdot\|_\infty$ or other polyedral gauges)



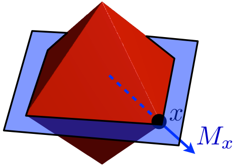
$M_x = \{z : \mathrm{supp}(z) = \mathrm{supp}(x)\}$

Structure of nonsmooth regularizers

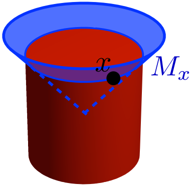
Many of the regularizers used in machine learning or image processing have a strong primal-dual structure (“mirror-stratifiable” [Fadili, M., Peyré ’18]) ...that can be exploit to get (enlarged) stability/identification results

Examples: (associated unit ball and low-dimensional manifold where x belongs)

- $R = \|\cdot\|_1$ (and $\|\cdot\|_\infty$ or other polyedral gauges)
- nuclear norm (aka trace-norm) $R(X) = \sum_i |\sigma_i(X)| = \|\sigma(X)\|_1$



$M_x = \{z : \mathrm{supp}(z) = \mathrm{supp}(x)\}$



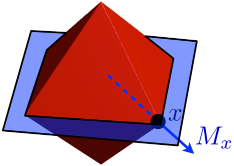
$M_x = \{z : \mathrm{rank}(z) = \mathrm{rank}(x)\}$

Structure of nonsmooth regularizers

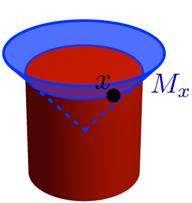
Many of the regularizers used in machine learning or image processing have a strong primal-dual structure ("mirror-stratifiable" [Fadili, M., Peyré '18]) ...that can be exploit to get (enlarged) stability/identification results

Examples: (associated unit ball and low-dimensional manifold where x belongs)

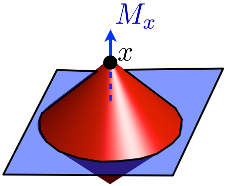
- $R = \|\cdot\|_1$ (and $\|\cdot\|_\infty$ or other polyedral gauges)
- nuclear norm (aka trace-norm) $R(X) = \sum_i |\sigma_i(X)| = \|\sigma(X)\|_1$
- group-ℓ_1 $R(x) = \sum_{b \in \mathcal{B}} \|x_b\|_2$ (e.g. $R(x) = \|x_{1,2}\| + |x_3|$)



$M_x = \{z : \mathrm{supp}(z) = \mathrm{supp}(x)\}$



$M_x = \{z : \mathrm{rank}(z) = \mathrm{rank}(x)\}$



$M_x = \{0\} \times \{0\} \times \mathbb{R}$

Recall on stratifications

A stratification of a set $D \subset \mathbb{R}^d$ is a (finite) partition $\mathcal{M} = \{M_i\}_{i \in I}$

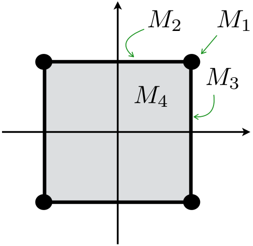
$$D = \bigcup_{i \in I} M_i$$

with so-called "strata" (e.g. smooth/affine manifolds) which fit nicely:

$$M \cap \mathrm{cl}(M') \neq \emptyset \quad \Longrightarrow \quad M \subset \mathrm{cl}(M')$$

Example: $\mathbb{B}_\infty$ the unit ℓ_∞-ball in $\mathbb{R}^2$

a stratification with 9 (affine) strata



Other examples: "tame" sets, remind Edouard's talk

Recall on stratifications

A stratification of a set $D \subset \mathbb{R}^d$ is a (finite) partition $\mathcal{M} = \{M_i\}_{i \in I}$

$$D = \bigcup_{i \in I} M_i$$

with so-called "strata" (e.g. smooth/affine manifolds) which fit nicely:

$$M \cap \mathrm{cl}(M') \neq \emptyset \quad \Longrightarrow \quad M \subset \mathrm{cl}(M')$$

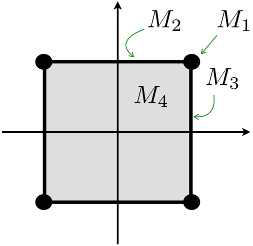
This relation induces a (partial) ordering $M \leqslant M'$

Example: $\mathbb{B}_\infty$ the unit ℓ_∞-ball in $\mathbb{R}^2$

a stratification with 9 (affine) strata

$$M_1 \leqslant M_2 \leqslant M_4$$

$$M_1 \leqslant M_3 \leqslant M_4$$



Other examples: "tame" sets, remind Edouard's talk

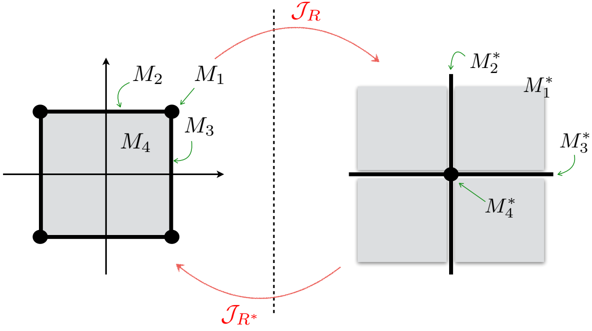
Mirror-stratifiable regularizations

(primal) stratification $\mathcal{M} = \{M_i\}_{i \in I}$ and (dual) stratification $\mathcal{M}^* = \{M_i^*\}_{i \in I}$

in one-to-one decreasing correspondence

through the transfert operator $\mathcal{J}_R(S) = \bigcup_{x \in S} \mathrm{ri}(\partial R(x))$

Simple example: $R = \iota_{\mathbb{B}_\infty}$ $R^* = \|\cdot\|_1$



$$\mathcal{J}_R(M_i) = \bigcup_{x \in M_i} \mathrm{ri}\, \partial R(x) = \mathrm{ri}\, N_{\mathbb{B}_\infty}(x) = M_i^* \qquad M_i = \mathrm{ri}\, \partial \|x\|_1 = \bigcup_{x \in M_i^*} \mathrm{ri}\, \partial R^*(x) = \mathcal{J}_{R^*}(M_i^*)$$

Enlarged stability result

Theorem (Fadili, M., Peyré '18)

For the composite optimization problem *(smooth + nonsmooth)*

$$\min_{x \in \mathbb{R}^d} \quad F(x, p) + R(x),$$

satisfying mild assumptions *(unique minimizer $x^\star(p_0)$ at p_0 and objective uniformly level-bounded in x)*, *if R is mirror-stratifiable, then for $p \sim p_0$,*

$$M_{x^\star(p_0)} \leqslant M_{x^\star(p)} \leqslant \mathcal{J}_{R^*}(M^*_{u^\star(p_0)})$$

If $R = \|\cdot\|_1$, then $\quad \operatorname{supp}(x^\star(p_0)) \subseteq \operatorname{supp}(x^\star(p)) \subseteq \{i : |u^\star(p_0)_i| = 1\}$

Enlarged stability result

Theorem (Fadili, M., Peyré '18)

For the composite optimization problem (smooth + nonsmooth)

$$\min_{x\in\mathbb{R}^d} \quad F(x,p) + R(x),$$

satisfying mild assumptions (unique minimizer $x^\star(p_0)$ at p_0 and objective uniformly level-bounded in x), *if R is mirror-stratifiable, then for $p \sim p_0$,*

$$M_{x^\star(p_0)} \leqslant M_{x^\star(p)} \leqslant \mathcal{J}_{R^*}(M^*_{u^\star(p_0)})$$

If $R = \|\cdot\|_1$, then $\quad \operatorname{supp}(x^\star(p_0)) \subseteq \operatorname{supp}(x^\star(p)) \subseteq \{i : |u^\star(p_0)_i| = 1\}$

Remark: Optimality conditions for a primal-dual solution $(x^\star(p), u^\star(p))$

$$u^\star(p) = -\nabla F(x^\star(p), p) \;\in\; \partial R(x^\star(p))$$

In the non-degenerate case: $\quad u^\star(p_0) \in \operatorname{ri}\left(\partial R(x^\star(p_0))\right)$

$$M_{x^\star(p_0)} = M_{x^\star(p)} \quad \left(= \mathcal{J}_{R^*}(M^*_{u^\star(p_0)})\right)$$

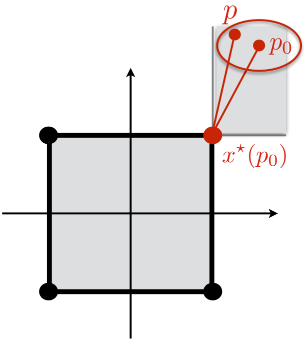
we have the exact stability, expected [Lewis '02]

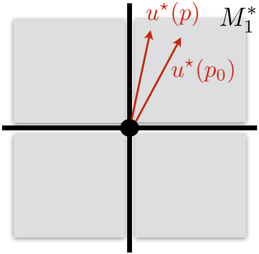
Enlarged stability illustrated

$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^n \end{cases}$$

Non-degenerate case: $u^\star(p_0) = p_0 - x^\star(p_0) \in \mathrm{ri}\ N_{\mathbb{B}_\infty}(x^\star(p_0))$

$\Longrightarrow M_1 = M_{x^\star(p_0)} = M_{x^\star(p)}$ (in this case $x^\star(p) = x^\star(p_0)$)

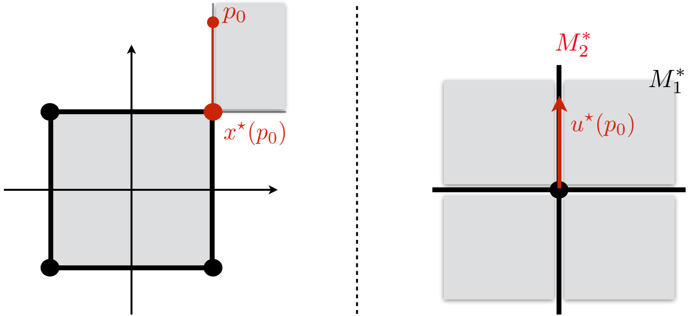




Enlarged stability illustrated

$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad\qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^n \end{cases}$$

General case: $u^\star(p_0) = p_0 - x^\star(p_0) \;\in\; \cancel{\mathrm{ri}}\, N_{\mathbb{B}_\infty}(x^\star(p))$

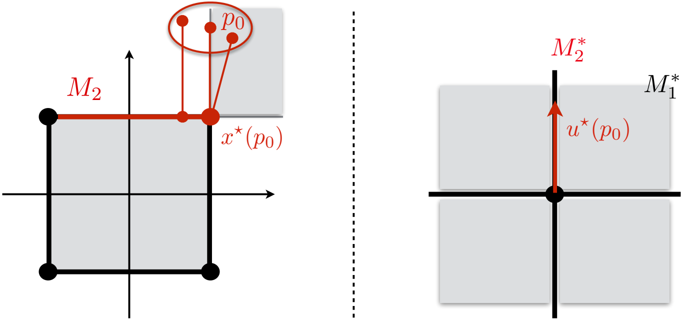


Enlarged stability illustrated

$$\left\{\begin{array}{ll} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{array}\right. \qquad\qquad \left\{\begin{array}{ll} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^n \end{array}\right.$$

General case: $u^\star(p_0) = p_0 - x^\star(p_0) \;\in\; \cancel{\mathrm{ri}}\, N_{\mathbb{B}_\infty}(x^\star(p))$

$$\Longrightarrow \; M_1 = M_{x^\star(p_0)} \leqslant M_{x^\star(p)} \leqslant \mathcal{J}_{R^*}(M^*_{u^\star(p_0)}) = M_2$$



Outline

Activity identification

Composite optimization problem (smooth + nonsmooth)

$$\min_{x\in\mathbb{R}^d} \quad F(x) + R(x)$$

Basic proximal-gradient algorithm

$$x_{k+1} = \operatorname{prox}_{\gamma R}\big(x_k - \gamma\nabla F(x_k)\big)$$

$$\operatorname{prox}_{\gamma R}(x) \;=\; \operatorname*{argmin}_y \; R(y) + \frac{1}{2\gamma}\|y - x\|^2$$

$\operatorname{prox}_{\gamma R}(x)$ easy to compute in some important cases

e.g. explicit expression for $R = \|\cdot\|_1$ (soft-thresholding)

Activity identification

Composite optimization problem (smooth + nonsmooth)

$$\min_{x \in \mathbb{R}^d} \; F(x) + R(x)$$

Basic proximal-gradient algorithm

$$x_{k+1} = \operatorname{prox}_{\gamma R}\big(x_k - \gamma \nabla F(x_k)\big)$$

$$\operatorname{prox}_{\gamma R}(x) \;=\; \operatorname*{argmin}_{y} \; R(y) + \frac{1}{2\gamma}\|y - x\|^2$$

$\operatorname{prox}_{\gamma R}(x)$ easy to compute in some important cases

e.g. explicit expression for $R = \|\cdot\|_1$ (soft-thresholding)

Identification: Beyond convergence

$\boxed{\text{after a finite moment } K\text{, all iterates } x_k \ (k \geqslant K) \text{ lie in an active set } M}$

- Used in e.g. safe screening [El Gahoui '12] [Salmon *et al* '19] [Sun *et al* '20]
- We even have bounds on K [Sun *et al* '19]
- When the problem is well-posed e.g. [Wright '96], [Lewis Drusvyatskiy '13]

Enlarged activity identification

Theorem (Fadili, M., Peyré '18)

Under convergence assumptions, if R is mirror-stratifiable, then for $k \geqslant K$

$$M_{x^\star} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}(M^*_{-\nabla F(x^\star)})$$

- Optimality condition $\quad -\nabla F(x^\star) \in \partial R(x^\star)$

  In the non-degenerate case: $\quad -\nabla F(x^\star) \in \mathrm{ri}\left(\partial R(x^\star)\right)$

  we have exact identification $M_{x^\star} = M_{x_k} \;\left(= \mathcal{J}_{R^*}(M^*_{-\nabla F(x^\star)})\right)$ [Liang *et al* 15]

Enlarged activity identification

Theorem (Fadili, M., Peyré '18)

Under convergence assumptions, if R is mirror-stratifiable, then for $k \geqslant K$

$$M_{x^\star} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}(M^*_{-\nabla F(x^\star)})$$

- Optimality condition $\quad -\nabla F(x^\star) \in \partial R(x^\star)$

  In the non-degenerate case: $\qquad -\nabla F(x^\star) \in \mathrm{ri}\left(\partial R(x^\star)\right)$

  we have exact identification $M_{x^\star} = M_{x_k} \; \left(= \mathcal{J}_{R^*}(M^*_{-\nabla F(x^\star)})\right)$ [Liang *et al* 15]

- In the general case: δ quantifies the degeneracy of the problem

$$\delta = \dim(\mathcal{J}_{R^*}(M^*_{-\nabla F(x^\star)})) - \dim(M_{x^\star})$$

  $\delta = 0$: well-posedness (fast convergence and identification)

  δ large : strong degeneracy (slow convergence and identification)

- Note: δ and K are not computable beforehand in general...

Illustration with nuclear norm

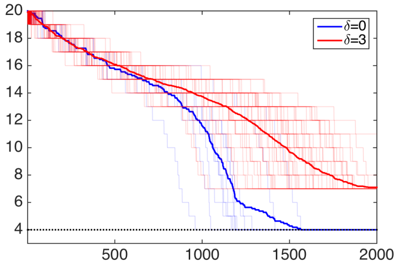
Matrix least-squares regularized by nuclear norm $\|X\|_* = \|\sigma(X)\|_1$

$$\min_{X\in\mathbb{R}^{d=m\times m}} \quad \frac{1}{2}\|A(X)-y\|^2 \ + \ \lambda\|X\|_*$$

Generate many random problems (with $m = 20$ and $n = 300$), solve them

Select those with $\mathrm{rank}(X^\star)=4$ and $\delta = 0$ or 3 $\quad (\delta = \#\{i : |\sigma_i(U^\star)| = 1\} - \mathrm{rank}(X^\star))$

Plot the decrease of $\mathrm{rank}(X_k)$ with $X_{k+1} = \mathrm{prox}_{\gamma\|\cdot\|_*}\left(X_k - \gamma A^*(A(X_k)-y)\right)$

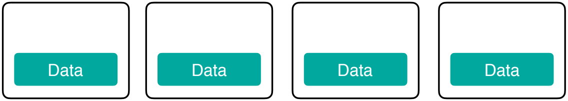


$\delta = 0$: well-posed vs. $\delta = 3$: degenerate

Outline

Basic distributed learning set-up

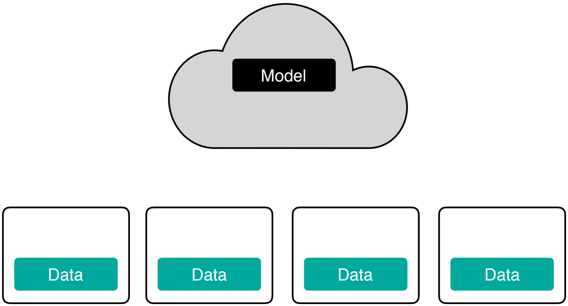
(Standard) centralized learning



- Data $(a_j, y_j)_{j=1,\ldots,n}$, prediction function $h(\cdot, x)$, model parameters $x \in \mathbb{R}^d$

Basic distributed learning set-up

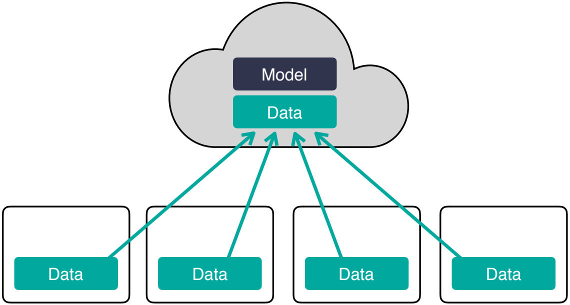
(Standard) centralized learning



- Data $(a_j, y_j)_{j=1,\ldots,n}$, prediction function $h(\cdot, x)$, model parameters $x \in \mathbb{R}^d$

Basic distributed learning set-up

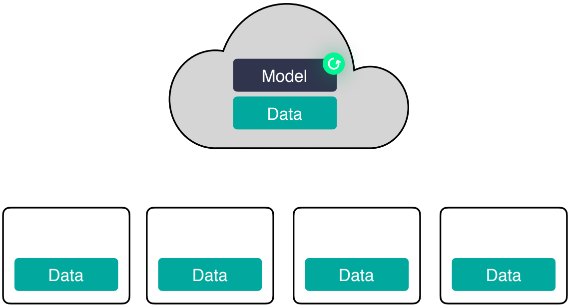
(Standard) centralized learning



- Data $(a_j, y_j)_{j=1,\ldots,n}$, prediction function $h(\cdot, x)$, model parameters $x \in \mathbb{R}^d$

Basic distributed learning set-up

(Standard) centralized learning



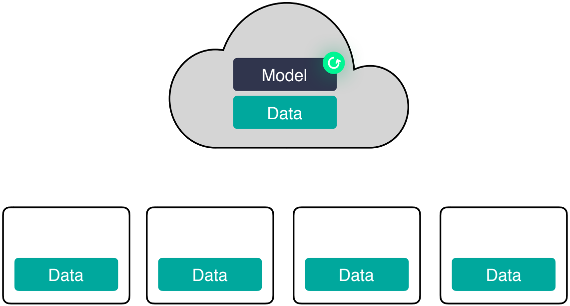
- Data $(a_j, y_j)_{j=1,\ldots,n}$, prediction function $h(\cdot, x)$, model parameters $x \in \mathbb{R}^d$
- Empirical risk minimization

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{j=1}^{n} \ell\big(y_j, h(a_j, x)\big) \quad + \quad \lambda R(x)$$

Basic distributed learning set-up

(Standard) centralized learning

- needs of lot of storage ☹
- is highly privacy invasive ☹

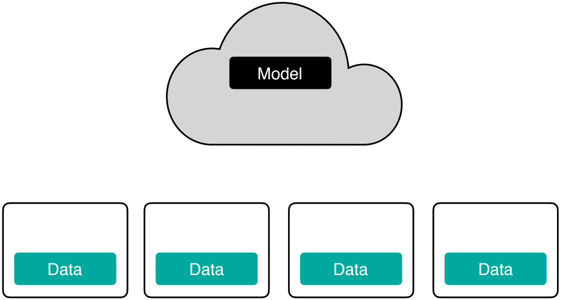


- Data $(a_j, y_j)_{j=1,\dots,n}$, prediction function $h(\cdot, x)$, model parameters $x \in \mathbb{R}^d$
- Empirical risk minimization

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{j=1}^{n} \ell\big(y_j, h(a_j, x)\big) \quad + \quad \lambda R(x)$$

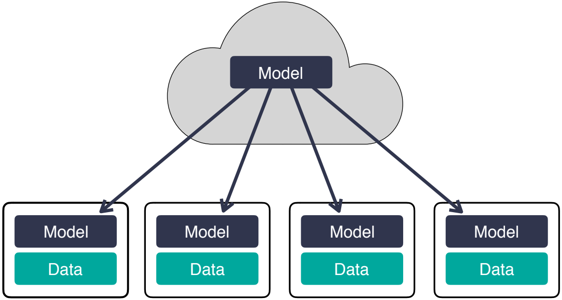
Move the model, not the data !

Collaborative/Federative learning (introduction of Aurélien's talk this morning)



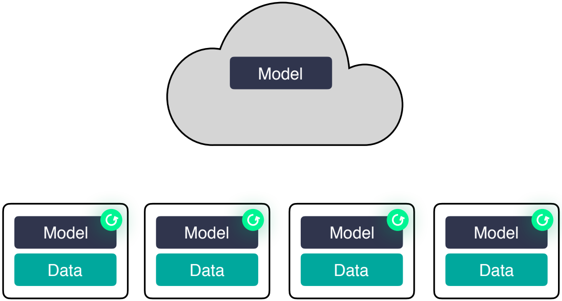
Move the model, not the data !

Collaborative/Federative learning (introduction of Aurélien's talk this morning)



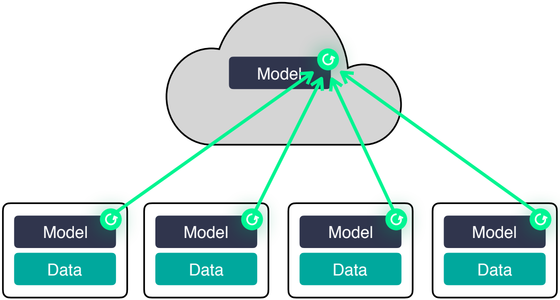
Move the model, not the data !

Collaborative/Federative learning (introduction of Aurélien's talk this morning)



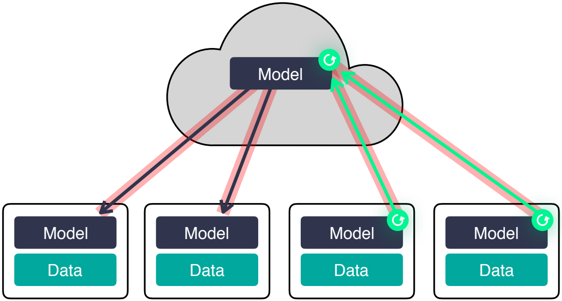
Move the model, not the data !

Collaborative/Federative learning (introduction of Aurélien's talk this morning)



Move the model, not the data !

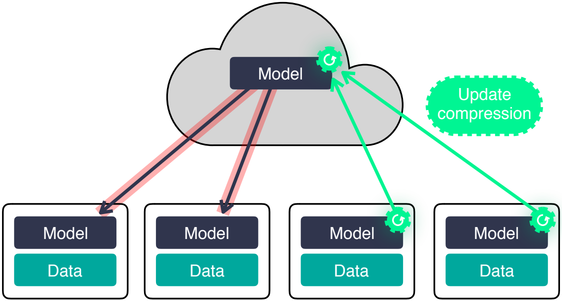
Collaborative/Federative learning (introduction of Aurélien's talk this morning)



- Communication is the bottleneck ☹
- We need compression ! Mikael talk, yesterday morning (?)

Move the model, not the data !

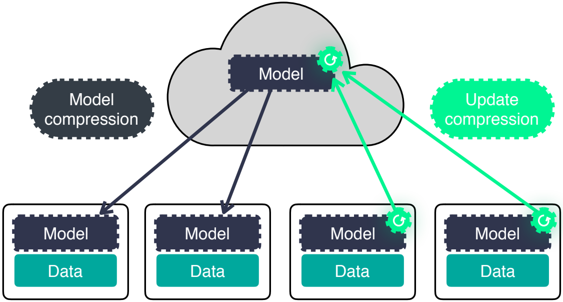
Collaborative/Federative learning (introduction of Aurélien's talk this morning)



- Communication is the bottleneck ☹
- We need compression ! Mikael talk, yesterday morning (?)

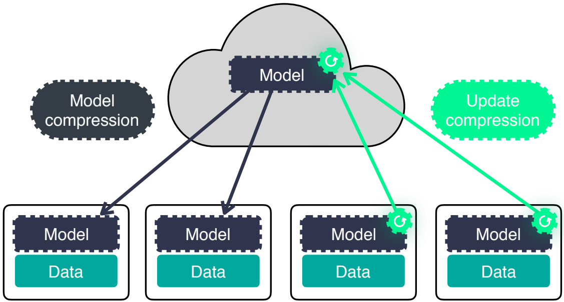
Move the model, not the data !

Collaborative/Federative learning (introduction of Aurélien's talk this morning)



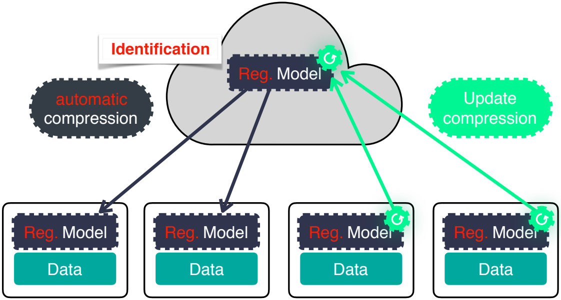
- Communication is the bottleneck ☹
- We need compression ! Mikael talk, yesterday morning (?)
- Many compression techniques... recall Martin's talk yesteday afternoon
- Let's discuss another one, complementary to existing ones

Application of identification to federated learning



Application of identification to federated learning

With nonsmooth regularizers, identification comes into play

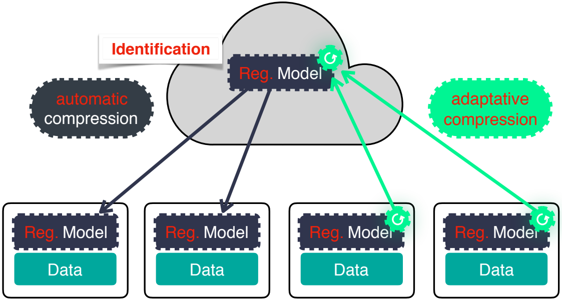


- Observation: identification gives automatic model compression

  e.g. for $R = \|\cdot\|_1$, model becomes sparse... just communicate nonzero entries!

Application of identification to federated learning

With nonsmooth regularizers, identification comes into play



- Observation: identification gives automatic model compression

  e.g. for $R = \|\cdot\|_1$, model becomes sparse... just communicate nonzero entries!

- [Grishchenko, Iutzeler, M. '19] uses again identification for update comp.

  $$\text{Project update onto} \quad M_{x_k} \quad + \quad \text{randomly selected } M$$

  e.g. for $R = \|\cdot\|_1$, select current support + random entries

- Algo with intricate convergence analysis due to non-uniform selection...

Illustration of communication-efficient proximal method

On an instance of TV-regularized logistic regression (a1a dataset on 10 machines)

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{j=1}^{n} \log\left(1+\exp(-y_j\langle a_j, x\rangle\right) \; + \; \lambda\,\mathrm{TV}(x)$$

Total Variation

$$\mathrm{TV}(x) = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

Comparison of
- Usual distributed proximal-gradient (black)
- Adaptive distributed proximal-subspace descent (red)

  for different selections M_{x_k} + random others

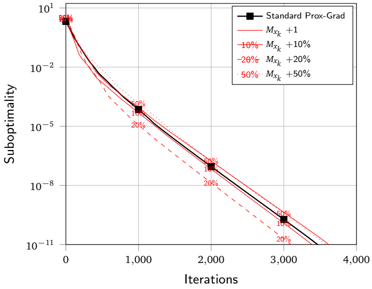


Illustration of communication-efficient proximal method

On an instance of TV-regularized logistic regression (a1a dataset on 10 machines)

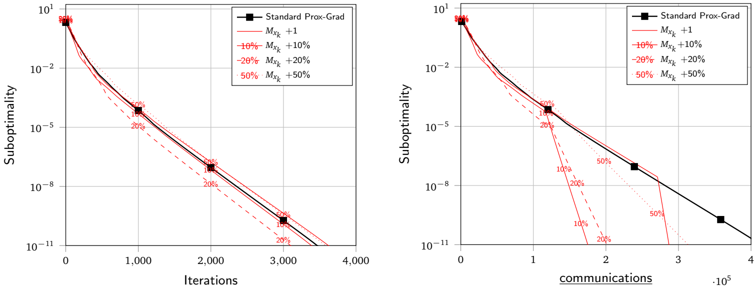
$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{j=1}^{n} \log\left(1+\exp(-y_j\langle a_j, x\rangle\right) \;+\; \lambda\, \mathrm{TV}(x)$$

Total Variation

$$\mathrm{TV}(x) = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

Comparison of
- Usual distributed proximal-gradient (black)
- Adaptive distributed proximal-subspace descent (red)

  for different selections M_{x_k} + random others



Acceleration... with respect to size of communication

Illustration of communication-efficient proximal method

On an instance of TV-regularized logistic regression (a1a dataset on 10 machines)

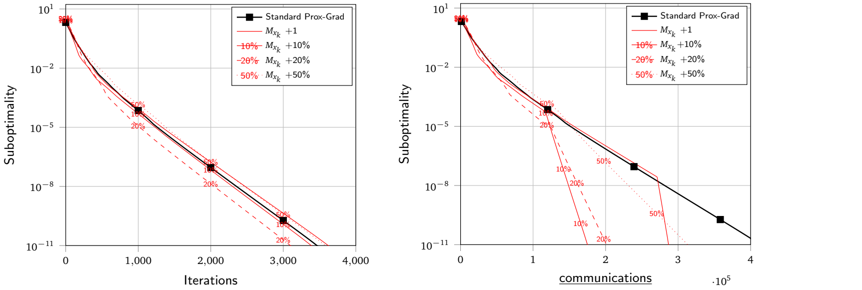
$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{j=1}^{n} \log\left(1+\exp(-y_j\langle a_j, x\rangle\right) \;+\; \lambda\, \mathrm{TV}(x)$$

Total Variation

$$\mathrm{TV}(x) = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

Comparison of
- Usual distributed proximal-gradient (black)
- Adaptive distributed proximal-subspace descent (red)

  for different selections M_{x_k} + random others



Acceleration... with respect to size of communication

Tradeoff between compression (less comm.) and identification (faster cv)

Outline

Supervised learning: model consistency ?

- Assume data $(a_i, y_i)_{i=1,\ldots,n}$ are sampled from linear model

  $y = \langle a, \bar{x} \rangle + \nu$ with random (a, ν)

- Structural assumption: $\bar{x}$ has a low-complexity for R

  $\bar{x} = \mathrm{argmin}_{x \in \mathbb{R}^d} \left\{ R(x) : x \in \mathrm{argmin}_{z \in \mathbb{R}^d} \mathbb{E}\left[(\langle a, z \rangle - y)^2 \right] \right\}$

- Regularized least-squares (if $R = \|\cdot\|_1$, this is LASSO)

$$\min_{x \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} \left(\langle a_i, x \rangle - y_i \right)^2 + \lambda_n R(x)$$

- Stochastic (proximal-)gradient algorithms (at iteration k, pick randomly $i(k)$)

$$x_{k+1} = \mathrm{prox}_{\gamma_k \lambda_n R} \left(x_k - \gamma_k \left(\left(\langle a_{i(k)}, x_k \rangle - y_{i(k)} \right) a_{i(k)} + \varepsilon_k \right) \right)$$

  E.g. SGD, SAGA [Delfazio *et al* '14], SVRG [Xiao-Zhang '14]

- Do we have model recovery/consistency i.e. $\boxed{x_k \in M_{\bar{x}} \text{ ?}}$

  (if we have enough observations, i.e. when $n \to +\infty$)

Enlarged identification of stochastic algorithms

Theorem (Garrigos, Fadili, M., Peyré '19)

Take $\lambda_n \to 0$ *with* $\lambda_n \sqrt{n/(\log\log n)} \to +\infty$. *If* n *large enough and for*

$$x_{k+1} = \mathrm{prox}_{\gamma_k \lambda_n R}\left(x_k - \gamma_k \big((\langle a_{i(k)}, x_k\rangle - y_{i(k)})\, a_{i(k)} + \varepsilon_k \big) \right)$$

with mild assumptions on errors ε_k *and stepsizes* γ_k. *Then, for* k *large, a.s.*

$$M_{\bar{x}} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}(M^*_{\bar{\eta}})$$

with $\quad \bar{\eta} = \underset{\eta\in\mathbb{R}^p}{\mathrm{argmin}} \left\{ \eta^\top C^\dagger \eta \,:\, \eta \in \partial R(\bar{x}) \cap \mathrm{Im}\, C \right\}$ *and* $C = \mathbb{E}\left[a a^\top \right]$

Comments:

- key dual object $\bar{\eta} \in \partial R(\bar{x})$ [Vaiter *et al* '16]
- λ_n decreases to 0, but not too fast
- SAGA and SVRG satisfy the "mild" assumption [Poon *et al* '18]
- (Prox-)SGD does not – and does not identify (e.g. [Lee Wright '12])

Enlarged identification of stochastic algorithms

Theorem (Garrigos, Fadili, M., Peyré '19)

Take $\lambda_n \to 0$ *with* $\lambda_n \sqrt{n/(\log\log n)} \to +\infty$. *If* n *large enough and for*

$$x_{k+1} = \mathrm{prox}_{\gamma_k \lambda_n R}\left(x_k - \gamma_k \big((\langle a_{i(k)}, x_k\rangle - y_{i(k)})\, a_{i(k)} + \varepsilon_k \big) \right)$$

with mild assumptions on errors ε_k *and stepsizes* γ_k. *Then, for* k *large, a.s.*

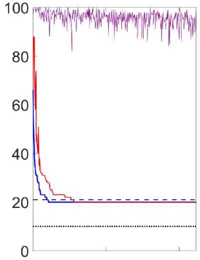
$$M_{\bar{x}} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}(M^*_{\bar{\eta}})$$

with $\quad \bar{\eta} = \underset{\eta\in\mathbb{R}^p}{\mathrm{argmin}} \left\{ \eta^\top C^\dagger \eta \,:\, \eta \in \partial R(\bar{x}) \cap \mathrm{Im}\, C \right\}$ *and* $C = \mathbb{E}\left[a a^\top \right]$

Comments:

- key dual object $\bar{\eta} \in \partial R(\bar{x})$ [Vaiter *et al* '16]
- λ_n decreases to 0, but not too fast
- SAGA and SVRG satisfy the "mild" assumption [Poon *et al* '18]
- (Prox-)SGD does not – and does not identify (e.g. [Lee Wright '12])

(on a LASSO instance)



Conclusion

Take-home message: identification often holds... and can be used

- Enlarged identification results (explaining observed phenomena)
- Better understanding of optim. algos (beyond convergence)
- Sparsify communications by adaptative dimension reduction

Conclusion

Take-home message: identification often holds... and can be used

- Enlarged identification results (explaining observed phenomena)
- Better understanding of optim. algos (beyond convergence)
- Sparsify communications by adaptative dimension reduction

Extensions, on-going work

- Many possible refinements of sensitivity results
  other data fidelity terms, a priori control on strata dimension, explaining transition curves...
- Use identification to accelerate convergence
  interplay between identification and acceleration (PhD of Gilles Bareilles)
- Subspace descent algorithms generalizing coordinate descent
  "coordinate" descent for nonseparable functions $\longrightarrow$ Franck's talk tomorow

Conclusion

Take-home message: identification often holds... and can be used

- Enlarged identification results (explaining observed phenomena)
- Better understanding of optim. algos (beyond convergence)
- Sparsify communications by adaptative dimension reduction

Extensions, on-going work

- Many possible refinements of sensitivity results
  other data fidelity terms, a priori control on strata dimension, explaining transition curves...
- Use identification to accelerate convergence
  interplay between identification and acceleration (PhD of Gilles Bareilles)
- Subspace descent algorithms generalizing coordinate descent
  "coordinate" descent for nonseparable functions $\longrightarrow$ Franck's talk tomorow

thanks !!