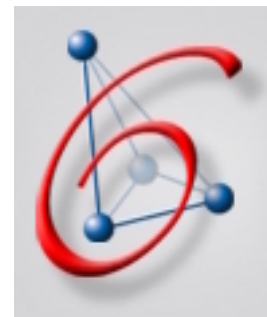
Analyse Non-Asymptotique d'un Compromis Computationellement Optimal pour les Méthodes Proximales Inexactes

École d'été de Peyresq

Pierre Machart

LIF, Aix-Marseille Université
LSIS, Université du Sud-Toulon-Var

http://www.lif.univ-mrs.fr/~pmachart/
pierre.machart@lif.univ-mrs.fr



26 juin

travaux effectués avec Sandrine Anthoine and Luca Baldassarre

Outline

Minimisation du risque

Cadre envisagé : l'apprentissage statistique supervisé

- Données : réalisations i.i.d. de v.a. $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y} \rightsquigarrow D$.
- But : apprendre un "bon" prédicteur $h : \mathcal{X} \to \mathcal{Y}$.

- Qualité d'une prédiction mesurée via une fonction de perte :

$$\ell : \mathcal{Y}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$$

- Qualité d'un prédicteur mesurée via une fonction de risque :

$$R(h) = \mathbb{E}_D \ell(h, \boldsymbol{x}, y)$$

Meilleur prédicteur possible :

$$h^* := \operatorname*{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$$

Limitations intrinsèques

Meilleur prédicteur possible :

$$h^* := \operatorname*{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} R(h).$$

Première limitation : $\mathcal{Y}^{\mathcal{X}}$ est trop difficile à explorer.

Solution : explorer une classe d'hypothèses (paramétrisée) $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$:

$$h^*_{\mathcal{H}} := \operatorname*{argmin}_{h \in \mathcal{H}} R(h).$$

Coût : induit une erreur d'approximation :

$$\mathcal{E}_{\text{app}} := R(h^*_{\mathcal{H}}) - R(h^*).$$

Limitations intrinsèques

Meilleur prédicteur dans $\mathcal{H}$:

$$h^*_{\mathcal{H}} := \operatorname*{argmin}_{h \in \mathcal{H}} R(h).$$

Seconde limitation : D n'est pas connue, de même que $R(h)$.

Solution : calculer un estimateur :

$$\widehat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, \boldsymbol{x}_i, y_i),$$

et le minimiser :

$$h_n := \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}(h) \big(+ \lambda \, \Omega(h) \big).$$

Coût : induit une erreur d'estimation :

$$\mathcal{E}_{\text{est}} := R(h_n) - R(h^*_{\mathcal{H}}).$$

Limitations intrinsèques

Estimateur du meilleur prédicteur dans $\mathcal{H}$:

$$h_n := \underset{h \in \mathcal{H}}{\operatorname{argmin}} \widehat{R}(h) \left(+\lambda \Omega(h)\right).$$

Troisième limitation : h_n n'a (en général) pas d'expression analytique.

Solution : faire une résolution numérique donnant $\tilde{h}_n$.

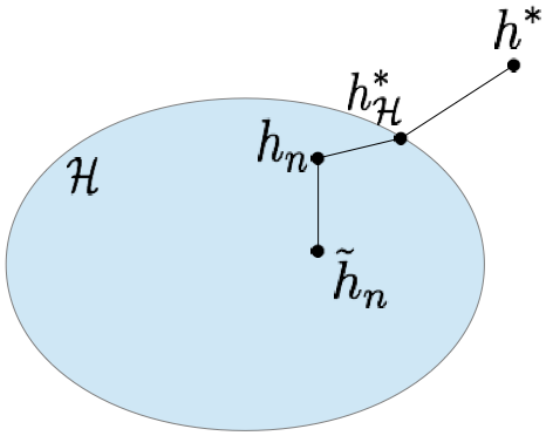
Coût : induit une erreur d'optimisation :

$$\mathcal{E}_{\text{opt}} := R(\tilde{h}_n) - R(h_n).$$

Décomposition de l'erreur

Les algos d'apprentissage fournissent un prédicteur $\tilde{h}_n$ avec une erreur :

$$\mathcal{E} := \mathcal{E}_{\text{app}} + \mathcal{E}_{\text{est}} + \mathcal{E}_{\text{opt}}.$$



réf : The Trade-Offs of Large-Scale Learning (Bottou et al., 2007)

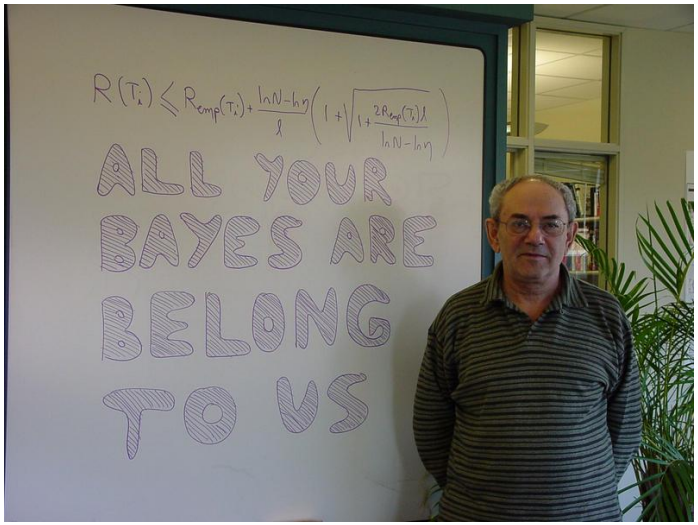
Décomposition de l'erreur

Les algos d'apprentissage fournissent un prédicteur $\tilde{h}_n$ avec une erreur :

$$\mathcal{E} := \mathcal{E}_{\text{app}} + \mathcal{E}_{\text{est}} + \cancel{\mathcal{E}_{\text{opt}}}.$$

Problème à petite échelle :

Statistiques



Vladimir Vapnik

Optimisation

Yurii Nesterov

réf : The Trade-Offs of Large-Scale Learning (Bottou et al., 2007)

Décomposition de l'erreur

Les algos d'apprentissage fournissent un prédicteur $\tilde{h}_n$ avec une erreur :

$$\mathcal{E} := \mathcal{E}_{\text{app}} + \mathcal{E}_{\text{est}} + \mathcal{E}_{\text{opt}}.$$

Problèmes à grande échelle :

Léon Bottou

réf : The Trade-Offs of Large-Scale Learning (Bottou et al., 2007)

Motivations

$$\mathcal{E} := \mathcal{E}_{\text{app}} + \mathcal{E}_{\text{est}} + \mathcal{E}_{\text{opt}}.$$

- L'efficacité computationelle est essentielle.
  $\Rightarrow$ Comment la mesurer ?
- Optimiser avec une précision finie.
  $\Rightarrow$ Les vitesses de convergence font-elles toujours sens ?
- Le temps de calcul est la ressource limitante.
  $\Rightarrow$ Comment la prendre en compte ?

Outline

Optimisation convexe non-différentiable

Problème général :

Minimisation d'une fonction composite

$$\min_x f(x) := g(x) + h(x),$$

avec $g : \mathbb{R}^n \to \mathbb{R}$ convexe, différentiable, avec gradient continue L-Lipschitz et $h : \mathbb{R}^n \to \mathbb{R}$ semi-continue inférieurement propre convexe.

Cadre général :

Méthodes par Gradient-Proximal :

Algorithm 1 Algorithme Proximal Exact

Require: initialisation x_0

for $i = 1$ à k **do**

$x_{i-\frac{1}{2}} = x_{i-1} - \frac{1}{L}\nabla g(x_{i-1})$ étape de descente de gradient

$x_i = \text{prox}_{h/L}(x_{i-\frac{1}{2}})$

end for

Méthodes proximales inexactes

Choix pour h :

- régularisation L_1, indicatrice sur un convexe...
  $\Rightarrow$ opérateur proximal calculable en forme fermée.
- régularisation TV, normes induisant de la parcimonie structurée...
  $\Rightarrow$ pas de solution analytique.

Algorithm 2 Algorithme Proximal Inexact

Require: initialisation x_0
  for $i = 1$ à k **do**
    $x_{i-\frac{1}{2}} = x_{i-1} - \frac{1}{L}\nabla g(x_{i-1})$ étape de descente de gradient
    while précision trop faible **do**
      Augmenter la précision de $\text{prox}_{h/L}(x_{i-\frac{1}{2}})$
    end while
    $x_i = \text{prox}_{h/L}(x_{i-\frac{1}{2}})$
  end for

$\Rightarrow$ Comment choisir la précision ?

Aperçu de la contribution

Coût global de la procédure d'optimisation :

$$C_{\text{glob}}(k, \{l_i\}_{i=1}^k) = C_{\text{in}} \sum_{i=1}^{k} l_i + kC_{\text{out}}.$$

La stratégie la plus rapide peut être obtenue en résolvant le problème d'optimisation suivant :

$$\min_{k, \{l_i\}_{i=1}^k} C_{\text{in}} \sum_{i=1}^{k} l_i + kC_{\text{out}} \qquad \text{s.t. } f(x_k) - f(x^*) \leq \rho.$$

Proposition (Stratégie optimale, (Machart et al., 2012b))

La stratégie la plus rapide est donnée par :

$$\forall\ i, l_i^* = cste,\ \ avec\ k^* = \operatorname*{argmin}_{k \in \mathbb{N}^*} fct(k).$$

Vitesses de convergence des méthodes proximales inexactes

Résolution numérique pour le calcul de chaque point proximal :

$$\frac{L}{2}\|x_k - z\|^2 + h(x_k) \leq \epsilon_k + \min_x \left\{ \frac{L}{2}\|x - z\|^2 + h(x) \right\}.$$

Algorithm 3 Boucle intérieure

while précision $< \epsilon_k$ **do**
  Augmenter la précision de $\text{prox}_{h/L}(x_{i-\frac{1}{2}})$
end while

Vitesses de convergence données par [Schmidt et al., 2011] :

$$f(x_k) - f(x^*) \leq \frac{L}{2k}\left(\|x_0 - x^*\| + 2\sum_{i=1}^{k} \sqrt{\frac{2\epsilon_i}{L}} + \sqrt{\sum_{i=1}^{k} \frac{2\epsilon_i}{L}} \right)^2.$$

$\Rightarrow$ Vitesses optimales si $\{\epsilon_k\}$ converge au moins en $O\left(\frac{1}{k^{(2+\delta)}}\right)$.

Mais cela impose un contrôle coûteux sur les approximations.

Précision et nombre d'itérations

Le point proximal est approximé par un algorithme itératif avec vitesse de convergence sous-linéaire :

$$\epsilon_i = \frac{A}{I_i^\alpha}.$$

Donne lieu à une borne paramétrée sur $f(x_k) - f(x^*)$:

$$f(x_k) - f(x^*) \leq B(k, \{I_i\}_{i=1}^k),$$

avec

$$B(k, \{I_i\}_{i=1}^k) = \frac{L}{2k}\left(\|x_0 - x^*\| + 3\sum_{i=1}^k \sqrt{\frac{2A}{LI_i^\alpha}}\right)^2.$$

Stratégie optimale

Soit $C(k) = \frac{\sqrt{L}}{3\sqrt{2A}} \left(\sqrt{\frac{2k\rho}{L}} - \|x_0 - x^*\| \right)$.

Proposition (Machart et al., 2012b)

Si $\rho < 6\sqrt{2LA}\|x_0 - x^*\|$, la solution de notre problème d'optimisation :

$$\min_{k, \{l_i\}_{i=1}^k} C_{in} \sum_{i=1}^{k} l_i + kC_{out} \qquad t.q.\ B(k, \{l_i\}_{i=1}^k) \leq \rho,$$

est :

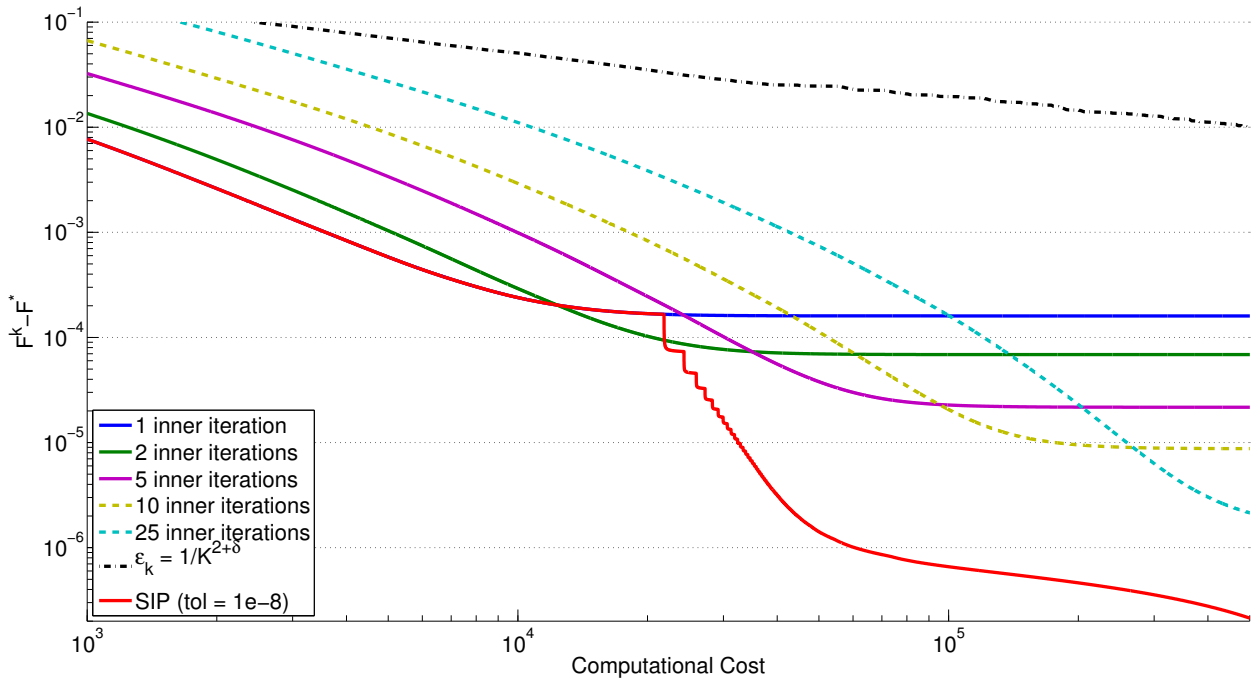
$$\forall\ i, l_i^* = \left(\frac{C(k^*)}{k^*} \right)^{-\frac{2}{\alpha}}, \ \textit{avec}\ k^* = \operatorname*{argmin}_{k \in \mathbb{N}^*} kC_{in} \left(\frac{C(k)}{k} \right)^{-\frac{2}{\alpha}} + kC_{out}.$$

Remarques :

- Nombre constant d'itérations intérieures (donc des ϵ_i).
- l_i^* tels que B vaut exactement ρ après k^* itérations extérieures.

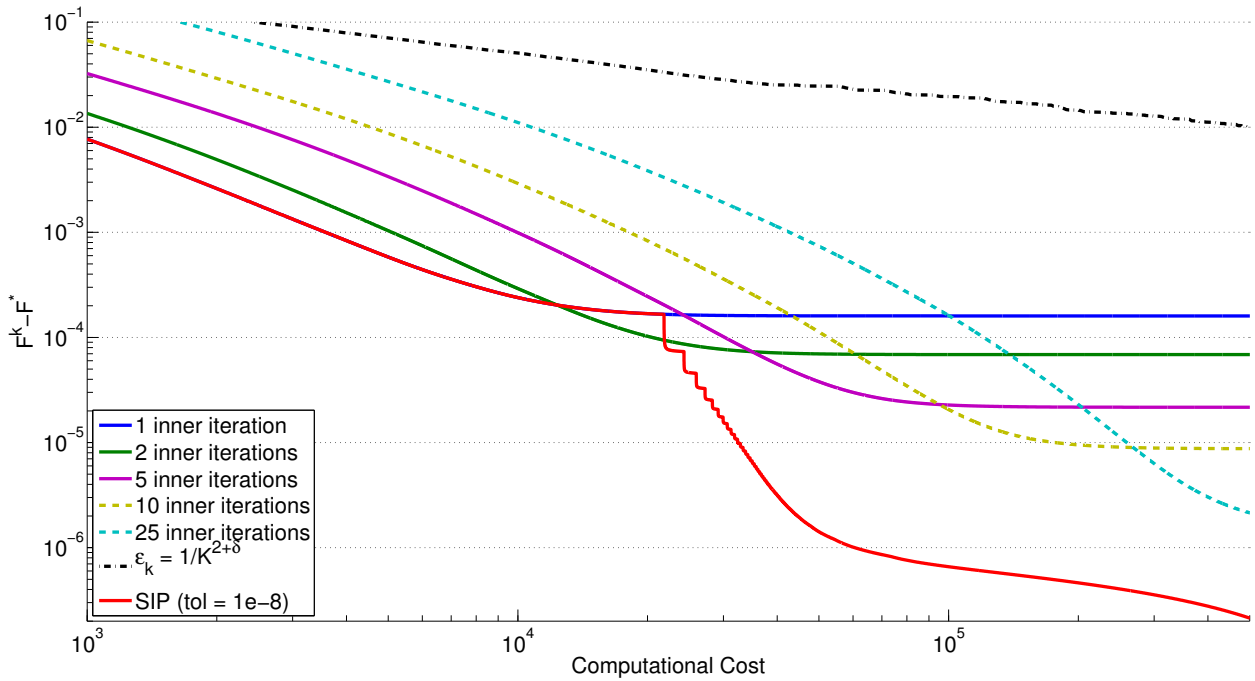
Simulations Numériques

Quelques simulations sur un problème de défloutage d'image.



Simulations Numériques

Quelques simulations sur un problème de défloutage d'image.



SIP (Speedy Inexact Proximal method) :

- stratégie adaptative
- d'excellentes performances en pratique

Outline

Conclusions et perspectives

Conclusions :

- Une nouvelle analyse en temps fini ($\neq$ analyses asymptotiques).
- Des stratégies optimales pour atteindre une solution à précision ρ.
- Une nouvelle stratégie SIP qui fonctionne très bien en pratique.

Perspectives :

- Mieux comprendre SIP.
- Nombres d'itérations constants $\Rightarrow$ régularisation ?
- Besoin de résultats optimistes en optimisation convexe.
- Peut-on appliquer la même méthodologie pour optimiser l'efficacité computationelle dans d'autres cadres ?

Qui a un chargeur de téléphone Nokia ?