Computational Statistics and Optimisation

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

Duality gap and stopping criterion

Back to gradient descent analysis

Forward-backward analysis

Forward-backward accelerated

Coordinate descent

Fenchel Duality for stopping criterion

F objective function, fix $\varepsilon > 0$ small, and stop when

$$\frac{F(\theta^{t+1}) - F(\theta_t)}{F(\theta^t)} \leqslant \varepsilon \text{ or } \nabla F(\theta^t) \leqslant \varepsilon$$

Alternative : leverage on the **duality gap**

Notation : $\boxed{F(\theta) = f(X\theta) + g(\theta)}$ with $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^p \to \mathbb{R}$ and X : $n \times p$ matrix.

Fenchel-Duality

Consider the problem $\min_\theta F(\theta)$, then the following holds

$$\sup_u \{-f^*(u) - g^*(-X^\top u)\} \leqslant \inf_\theta \{f(X\theta) + g(\theta)\}$$

Moreover, if f and g are **convex**, then under mild assumptions, equality of both sides holds (**strong duality**, no **duality gap**)

proof : use Fenchel-Young inequality (**TO DO: Blackboard**)

Fenchel Duality

We denote by

- $\theta^\star$: primal optimal solution of $\inf_\theta\{f(X\theta) + g(\theta)\}$
- u^* : dual solution of $\sup_u\{-f^*(u) - g^*(-X^\top u)\}$

Define the **duality gap** by :

$$\Delta(\theta, u) = F(\theta) + f^*(u) + g^*(-X^\top u)$$

Property of the duality gap

$$\forall \theta, u, \quad \Delta(\theta, u) \geqslant F(\theta) - F(\theta^\star) \geqslant 0$$

proof : Fenchel-duality applied to a primal solution $\theta^\star$

Motivation : $\boxed{\Delta(\theta, u) \leqslant \varepsilon \Rightarrow F(\theta) - F(\theta^\star) \leqslant \varepsilon}$

Example : Duality gap for the Lasso

Lasso objective : $\boxed{F(\theta) = \frac{1}{2}\|X\theta - y\|_2^2 + \lambda\|\theta\|_1}$

- $f(z) = \frac{1}{2}\|z - y\|_2^2; f^*(u) = \frac{1}{2}\|u\|_2^2 + \langle u, y\rangle$ (translation prop.)
- $g(\theta) = \lambda\|\theta\|_1; g^*(u) = \iota_{\{u, \|u\|_\infty \leqslant \lambda\}}$ (ℓ_∞ ball indicator)

- Duality gap : $$\begin{aligned}\Delta(\theta, u) =& F(\theta) + f^*(u) + g^*(-X^\top u)\\ =& F(\theta) + \frac{1}{2}\|u\|_2^2 + \langle u, y\rangle\end{aligned}$$

  as soon as $\|X^\top u\|_\infty \leqslant \lambda$, otherwise the bound is $+\infty$: useless

<u>Rem</u>: at optimum solutions and under mild assumptions
$\Delta(\theta^\star, u^*) = 0$

Example : Duality gap for the Lasso (II)

Possible choice :

- θ_t (current iterate of any iterative algorithm),
- $r_t = X\theta_k - y$ (minus current residuals)
- $u_t = \mu_t r_t$ with $\mu_t = \min(1, \lambda/\|X^\top r_t\|_\infty)$

Motivation for this choice : at optimum $u^* = \nabla f(X\theta^\star)$

Stopping criterion :

$$
\begin{aligned}
&\frac{1}{2}\|r_t\|_2^2 + \lambda\|\theta_t\|_1 + \frac{1}{2}\|u_t\|_2^2 + \langle u_t, y\rangle \leqslant \varepsilon \\
\Leftrightarrow\ &\frac{1}{2}(1+\mu_t^2)\|r_t\|_2^2 + \lambda\|\theta_t\|_1 + \mu_t\langle r_t, y\rangle \leqslant \varepsilon
\end{aligned}
$$

Plan

Duality gap and stopping criterion

Back to gradient descent analysis

Forward-backward analysis

Forward-backward accelerated

Coordinate descent

Convergence : Lipschitz gradient

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(x^t)}$$

Convergence rate for fixed step size

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta'), \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leqslant L \|\theta - \theta'\|$$

Result : for any minimum $\theta^\star$ of f, if $\alpha \leqslant \frac{1}{L}$ then θ^T satisfies

$$f(\theta^T) - f(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha T}$$

Rem: if f is twice differentiable $\nabla^2 f(x) \preceq L \cdot Id$

Example: $\theta \mapsto \frac{\|X\theta - y\|_2^2}{2}$ then $L = \lambda_{\max}(X^\top X)$ (spectral radius)

Convergence : proof

Point 1 : gradient L-Lipschitz implies quadratic upper bound

$$\forall (x, y) \quad f(x) \leqslant f(y) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Point 2 : use definition $\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)$ and get

$$f(\theta^{t+1}) \leqslant f(\theta^t) - (1 - \frac{L\alpha}{2})\alpha \|\nabla f(\theta^t)\|^2$$

Point 3 : use convexity, $0 < \alpha \leqslant \frac{1}{L}$ and $ab = (a^2 + b^2 - (a-b)^2)/2$

$$\begin{aligned} f(\theta^{t+1}) \leqslant & f(\theta^\star) + \nabla f(\theta^t)^\top (\theta^t - \theta^\star) - \frac{\alpha}{2}\|\nabla f(\theta^t)\|^2 \\ = & f(\theta^\star) + \frac{1}{2\alpha}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2) \end{aligned}$$

Convergence proof (bis)

Point 4 : Telescopic sums

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\big(f(\theta^{t+1})-f(\theta^\star)\big) &\leqslant \frac{1}{T}\frac{1}{2\alpha}(\|\theta^0-\theta^\star\|^2-\|x^T-\theta^\star\|^2)\\ &\leqslant \frac{1}{2\alpha T}\|\theta^0-\theta^\star\|^2\end{aligned}$$

From Point 2 $f(\theta^{t+1}) \leqslant f(\theta^t)$, hence

$$f(\theta^{t+1})-f(x^\star) \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\big(f(\theta^{t+1})-f(\theta^\star)\big) \leqslant \frac{1}{2\alpha T}\|\theta^0-\theta^\star\|^2$$

Plan

Duality gap and stopping criterion

Back to gradient descent analysis

Forward-backward analysis

Forward-backward accelerated

Coordinate descent

Composite minimization

One aims at minimizing :

$$F = f + g$$

- f smooth : ∇f is L-Lipschitz
- g proximable (prox-capable) :

$$\operatorname{prox}_{\alpha g}(y) = \arg\min_{z \in \mathbb{R}^d} \left(\frac{1}{2} \|z - y\|_2^2 + \alpha g(z) \right)$$

  "efficiently" computable

Examples: Projection over a box-constraint, (block) Soft-thresholding operator, shrinkage operator (Ridge)

Forward-Backward algorithm

Notation : $\boxed{\phi_\alpha(\theta) := \operatorname{prox}_{\alpha g}\left(\theta - \alpha \nabla f(\theta)\right)}$

Forward-Backward algorithm

Input: Initialization θ^0, step size α
Result: θ^T
while *not converged* **do**
  $\theta^{t+1} = \phi_\alpha(\theta^t)$
end

Rem:

$$\phi_\alpha(\theta) = \arg\min_{\theta'} \left(f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{1}{2\alpha} \|\theta' - \theta\|^2 + g(\theta') \right)$$

Convergence : f gradient Lipschitz

$$\boxed{\theta^{t+1} = \phi_\alpha(\theta^t) = \mathrm{prox}_{\alpha g}(\theta^t - \alpha \nabla f(\theta^t))}$$

Convergence rate for fixed step size

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta'), \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leqslant L \|\theta - \theta'\|$$

Result : for any minimum $\theta^\star$ of F, if $\alpha \leqslant \frac{1}{L}$ then θ^T satisfies

$$F(\theta^T) - F(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha T}$$

<u>Rem</u>: same bound as in the case with $g \equiv 0$

Proof : gradient

Point 1 : for $\alpha \leqslant 1/L$ and $\hat{x} = \phi_\alpha(\bar{x})$ then for all y :

$$\boxed{F(\hat{x}) + \frac{\|\hat{x} - y\|_2^2}{2\alpha} \leqslant F(y) + \frac{\|\bar{x} - y\|_2^2}{2\alpha}}$$

Proof : $H_\alpha(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + \frac{1}{2\alpha}\|y - \bar{x}\|^2 + g(y)$
H_α is $1/\alpha$-strongly convex since $\alpha \leqslant 1/L$ and $H(\cdot) - 1/(2\alpha)\|\cdot\|_2^2$ is convex (*cf.* for instance page 280, Hiriart-Urruty and Lemaréchal (1993))

$$\hat{x} = \arg\min_y H_\alpha(y) \quad \text{(\textit{cf.} two slides up)}$$

By $1/\alpha$-strong convexity : $\forall y, H_\alpha(\hat{x}) + 1/(2\alpha)\|\hat{x} - y\|_2^2 \leqslant H_\alpha(y)$

Point 1 (continued)

$$g(\hat{x}) + f(\bar{x}) + \langle \nabla f(\bar{x}), \hat{x} - \bar{x} \rangle + \frac{1}{2\alpha}(\|\hat{x} - \bar{x}\|^2 + \|\hat{x} - y\|^2) \leqslant$$
$$g(y) + f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + \frac{1}{2\alpha}\|y - \bar{x}\|^2$$

By convexity of f :

$$f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle \leqslant f(y)$$

and by the choice $\alpha \leqslant 1/L$ the following bound holds :

$$f(\hat{x}) \leqslant f(\bar{x}) + \langle \nabla f(\bar{x}), \hat{x} - \bar{x} \rangle + \frac{1}{2\alpha}\|\hat{x} - \bar{x}\|^2$$

Hence Point 1 : $\boxed{F(\hat{x}) + \frac{1}{2\alpha}\|\hat{x} - y\|^2 \leqslant F(y) + \frac{1}{2\alpha}\|y - \bar{x}\|^2}$

Theorem proof

Point 1 states : $F(\hat{x}) + \frac{1}{2\alpha}\|\hat{x} - y\|^2 \leqslant F(y) + \frac{1}{2\alpha}\|y - \bar{x}\|^2$, so choosing $y = \theta^\star$ (any minimizer of F) and $\bar{x} = \theta^t$, $\hat{x} = \theta^{t+1}$:

$$F(\theta^{t+1}) + \frac{1}{2\alpha}\|\theta^{t+1} - \theta^\star\|^2 \leqslant F(\theta^\star) + \frac{1}{2\alpha}\|\theta^\star - \theta^t\|^2$$

This leads to Point 3 of the smooth case :

$$F(\theta^{t+1}) \leqslant F(\theta^\star) + \frac{1}{2\alpha}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2)$$

and then the same telescopic argument lead to the bound

Plan

Duality gap and stopping criterion

Back to gradient descent analysis

Forward-backward analysis

Forward-backward accelerated

Coordinate descent

Forward-Backward accelerated algorithm

Notation : $\boxed{\phi_\alpha(\theta) := \mathrm{prox}_{\alpha g}\left(\theta - \alpha \nabla f(\theta)\right)}$

Forward-Backward algorithm

Input: Initialization θ^0, step size α, a sequence $(\mu_t)_{t\in\mathbb{N}}$ satisfying :$\mu_1 = 1$ and $\mu_{t+1}^2 - \mu_{t+1} \leqslant \mu_t^2$

Result: θ^T

while *not converged* **do**

$\theta^{t+1} = \phi_\alpha(z^t)$

$z^{t+1} = \theta^{t+1} + \frac{\mu_t - 1}{\mu_{t+1}}(\theta^{t+1} - \theta^t)$

end

Examples of admissible sequences :

- $\mu_{t+1} = \sqrt{\mu_t^2 + 1/4} + 1/2$ (*i.e.*, $\mu_{t+1}^2 - \mu_{t+1} = \mu_t^2$)
- $\mu_{t+1} = (t+1)/2$
- $\mu_{t+1} = (t + a - 1)/a$

Convergence : Lipschitz gradient

$$\boxed{\theta^{t+1} = \phi_\alpha(\theta^t) = \operatorname{prox}_{\alpha g}(\theta^t - \alpha \nabla f(\theta^t))}$$

Convergence rate for fixed step size

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta'), \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leqslant L \|\theta - \theta'\|$$

Result : for any minimum $\theta^\star$ of F, if $\alpha \leqslant \frac{1}{L}$ then θ^T satisfies

$$F(\theta^T) - F(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha\mu_T^2}$$

Rem: for common choices given above $\mu_t \approx t$

Rem: define $F^* = F(\theta^\star)$ for the proof

Proof : rate for the Nesterov acceleration

Point 1 with $\hat{x} = \phi_\alpha(z^t)$, $\bar{x} = z^t$, $y = (1 - 1/\mu_{t+1})\theta^t + 1/\mu_{t+1} \cdot \theta^\star$

$$\boxed{F(\hat{x}) + \frac{\|\hat{x} - y\|_2^2}{2\alpha} \leqslant F(y) + \frac{\|\bar{x} - y\|_2^2}{2\alpha}}$$

with $u^{t+1} = \theta^t + \mu_{t+1}(\theta^{t+1} - \theta^t)$ and a little algebra gives :

$$\begin{aligned}
F(\theta^{t+1}) + \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} &\leqslant F(y) + \frac{\|u^t - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} \\
F(\theta^{t+1}) - F^* - (1 - \frac{1}{\mu_{t+1}})(F(\theta^t) - F^*) &\leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} \\
\mu_{t+1}^2 \Delta F_{t+1}^* - (\mu_{t+1}^2 - \mu_{t+1})(\Delta F_t^*) &\leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha}
\end{aligned}$$

(convexity of F and $\Delta F_{t+1}^* = F(\theta^{t+1}) - F^*$)

Proof continued

Define $\rho_{t+1} := \mu_{t+1} - \mu_{t+1}^2 + \mu_t^2 \geqslant 0$ so

$$\mu_{t+1}^2 \Delta F_{t+1}^* - (\mu_{t+1}^2 - \mu_{t+1})(\Delta F_t^*) \leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha}$$

$$\mu_{t+1}^2 \Delta F_{t+1}^* - \mu_t^2 \Delta F_t^* + \rho_{t+1} \Delta F_t^* \leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha}$$

Telescopic terms again (convention $\mu_0 = 0$ and $u_0 = x_0 = x_{-1}$)

$$\mu_T^2 \Delta F_T^* + \sum_{t=0}^{T} \rho_{t+1} \Delta F_t^* \leqslant \frac{\|u^0 - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^T - \theta^\star\|_2^2}{2\alpha}$$

$$\mu_T^2 \Delta F_T^* \leqslant \frac{\|u^0 - \theta^\star\|_2^2}{2\alpha}$$

Convergence of the iterates

Very recent result : Chambolle and Dossal 2014 Proof out of the scope of this course
More reading on the previous theme :

- Nesterov for proofs, strong convexity, etc.
- Beck and Teboulle09 for ISTA/FISTA analysis
- Chambolle and Dossal 2014 for FISTA with larger choice of updating rules

Plan

Duality gap and stopping criterion

Back to gradient descent analysis

Forward-backward analysis

Forward-backward accelerated

Coordinate descent

Coordinate descent

Objective : solve $\arg\min_{\theta \in \mathbb{R}^p} f(\theta)$

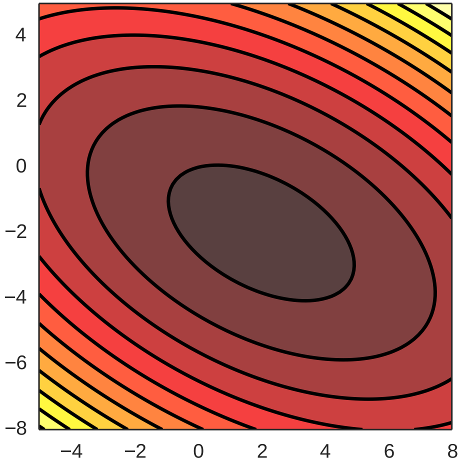
Initialization : $\theta^{(0)}$

$$
\begin{aligned}
\theta_1^{(k)} &\in \arg\min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)} \dots, \theta_p^{(k-1)}) \\
\theta_2^{(k)} &\in \arg\min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_p^{(k-1)}) \\
\theta_3^{(k)} &\in \arg\min_{\theta_3 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_p^{(k-1)}) \\
&\vdots \\
\theta_p^{(k)} &\in \arg\min_{\theta_p \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}, \dots, \theta_p)
\end{aligned}
$$

cycle over coordinates

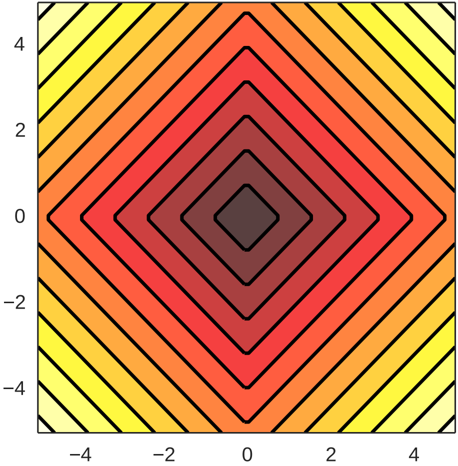
Motivation

- Convergence guarantees toward a minimum for smooth functions *cf.* Tseng (2001)



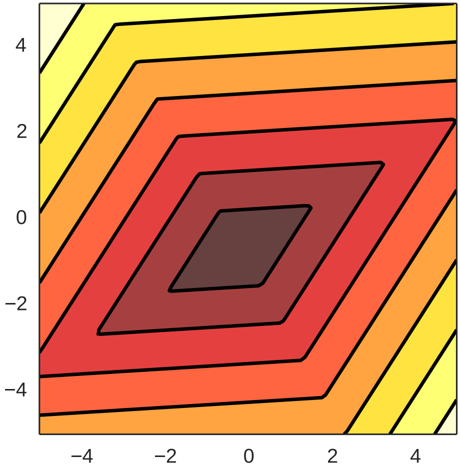
Motivation

- Convergence guarantees toward a minimum for separable functions *cf.* Tseng (2001)



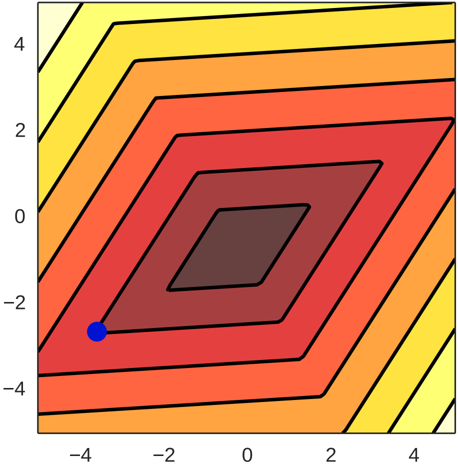
Motivation

- NO CONVERGENCE toward a minimum for non-separable/non-smooth functions : some points get stuck



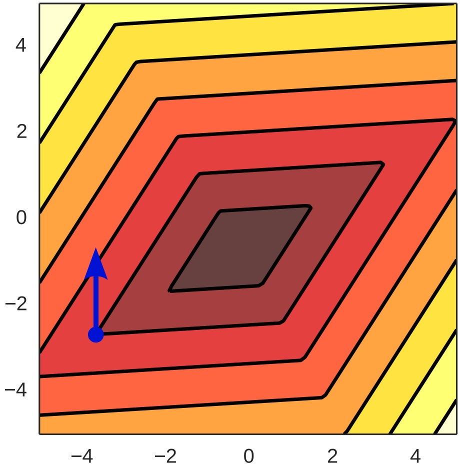
Motivation

- NO CONVERGENCE toward a minimum for non-separable/non-smooth functions : some points get stuck



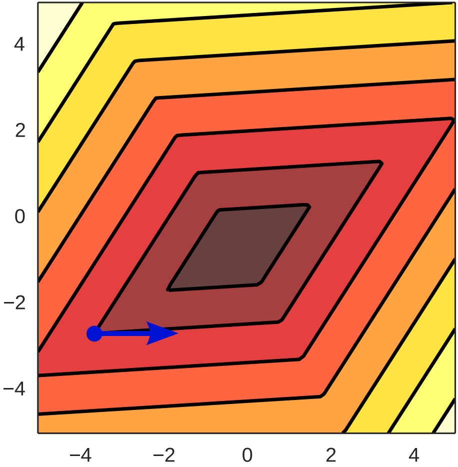
Motivation

- NO CONVERGENCE toward a minimum for non-separable/non-smooth functions : some points get stuck



Motivation

- NO CONVERGENCE toward a minimum for non-separable/non-smooth functions : some points get stuck



Motivation

- Coordinate descent can be extremely fast
- Possibly visit the coordinate in arbitrary order (cycle, random, more refined methods,etc.)
- Possibly by blocks : update not only one coordinate, but a bunch of them (optimize according to your architecture)

Exo: Testing over a the ridge regression problem the relative performance of cycling and random sampling of the coordinate

---

CD for least square II

$$\arg\min_{\theta\in\mathbb{R}^p} f(\theta) \text{ pour } f(\theta) = \frac{1}{2}\|y - X\theta\|_2^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\theta_j \mathbf{x}_j)^2$$

Reminder : $\nabla f(\theta) = X^\top (X\theta - y) = \begin{pmatrix} \mathbf{x}_1^\top (X\theta - y) \\ \vdots \\ \mathbf{x}_p^\top (X\theta - y) \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial f}{\partial \theta_p}(\theta) \end{pmatrix}$

Minimize w.r.t θ_j with fixed θ_k $(k \neq j)$

$$0 = \frac{\partial f}{\partial \theta_j}(\theta) = \mathbf{x}_j^\top (X\theta - y) = \mathbf{x}_j^\top \left(\mathbf{x}_j\theta_j + \sum_{k\neq j}\mathbf{x}_k\theta_k - y\right)$$

$$\Leftrightarrow \theta_j = \frac{\mathbf{x}_j^\top \left(y - \sum_{k\neq j}\mathbf{x}_k\theta_k\right)}{\mathbf{x}_j^\top \mathbf{x}_j} = \frac{\mathbf{x}_j^\top \left(y - \sum_{k=1}^{p}\mathbf{x}_k\theta_k + \mathbf{x}_j\theta_j\right)}{\|\mathbf{x}_j\|_2^2}$$

CD for least square II

Clever update scheme with low memory impact by storing :

- current **residual** in a variable $r^{(k)}$ (size n vector)
- current **coefficient** in $\theta^{(k)}$ (size p vector)

Coordinate descent for least square

Input: Observations y, features $X = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$, initial $\theta^{(0)}$
Result: Vector $\theta^{(K)}$
while *not converged* **do**

Pick a coordinate j

$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$

$\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2$

$r^{(k+1)} = r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$

end

Rem: computational simplification $\|\mathbf{x}_j\|_2^2 = 1$ (NORMALIZE !)
Rem: the residual update can be done in place

Ridge regression with coordinate descent

$$\arg\min_{\theta\in\mathbb{R}^p} f(\theta) \text{ for } f(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\theta_j \mathbf{x}_j)^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\theta_j^2$$

$$\nabla f(\theta) = X^\top (X\theta - y) + \lambda\theta = \begin{pmatrix} \mathbf{x}_1^\top (X\theta - y) + \lambda\theta_1 \\ \vdots \\ \mathbf{x}_p^\top (X\theta - y) + \lambda\theta_p \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial\theta_1}(\theta) \\ \vdots \\ \frac{\partial f}{\partial\theta_p}(\theta) \end{pmatrix}$$

Minimize w.r.t θ_j fixing θ_k $(k \neq j)$

$$0 = \frac{\partial f}{\partial\theta_j}(\theta) = \mathbf{x}_j^\top (X\theta - y) + \lambda\theta_j = \mathbf{x}_j^\top \left(\mathbf{x}_j\theta_j + \sum_{k\neq j}\mathbf{x}_k\theta_k - y\right) + \lambda\theta_j$$

$$\Leftrightarrow \theta_j = \frac{\mathbf{x}_j^\top \left(y - \sum_{k\neq j}\mathbf{x}_k\theta_k\right)}{\mathbf{x}_j^\top\mathbf{x}_j + \lambda} = \frac{\mathbf{x}_j^\top \left(y - \sum_{k=1}^{p}\mathbf{x}_k\theta_k + \mathbf{x}_j\theta_j\right)}{\|\mathbf{x}_j\|_2^2 + \lambda}$$

Ridge regression with coordinate descent II

Clever update scheme with low memory impact by storing :

- current **residual** in a variable $r^{(k)}$ (size n vector)
- current **coefficient** in $\theta^{(k)}$ (size p vector)

Coordinate Descent for Ridge regression

Input: Observations y, features $X = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$, initial $\theta^{(0)}$
Result: Vector $\theta^{(K)}$
while *not converged* **do**
  Pick a coordinate j
  $r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$
  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda)$
  $r^{(k+1)} = r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$
end

Rem: computational simplification $\|\mathbf{x}_j\|_2^2 = 1$ (NORMALIZE !)
Rem: the residual update can be done in place

Lasso with coordinate descent

$$\arg\min_{\theta \in \mathbb{R}^p} f(\theta) \text{ for } f(\theta) = \frac{1}{2}\|y - X\theta\|^2 + \lambda \sum_{j=1}^{p} |\theta_j|$$

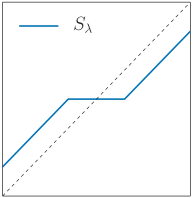
Minimize w.r.t θ_j fixing θ_k $(k \neq j)$

$$\begin{aligned}
\hat{\theta}_j &= \arg\min_{\theta_j \in \mathbb{R}} f(\theta_1, \cdots, \theta_p) \\
&= \arg\min_{\theta_j \in \mathbb{R}} \frac{1}{2}\|y - \sum_{k \neq j} \theta_k \mathbf{x}_k - \mathbf{x}_j \theta_j\|^2 + \lambda \sum_{k \neq j} |\theta_j| + \lambda |\theta_j| \\
&= \arg\min_{\theta_j \in \mathbb{R}} \frac{1}{2}\|\mathbf{x}_j\|^2 \theta_j^2 - \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \theta_j + \lambda |\theta_j| \\
&= \arg\min_{\theta_j \in \mathbb{R}} \|\mathbf{x}_j\|^2 \left[\frac{1}{2} \left(\theta_j - \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)^2 + \frac{\lambda}{\|\mathbf{x}_j\|^2} |\theta_j| \right]
\end{aligned}$$

Lasso with coordinate descent

$$\hat{\theta}_j = \underset{\theta_j \in \mathbb{R}}{\arg\min} \|\mathbf{x}_j\|^2 \left[\frac{1}{2} \left(\theta_j - \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)^2 + \frac{\lambda}{\|\mathbf{x}_j\|^2} |\theta_j| \right]$$

Soft-Thresholding : $\quad S_\lambda(z) = \underset{x \in \mathbb{R}}{\arg\min}\, x \mapsto \frac{1}{2}(z - x)_2^2 + \lambda |x|$



Update rule : $\quad \hat{\theta}_j = S_{\lambda / \|\mathbf{x}_j\|^2} \left(\|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)$

Ridge regression with coordinate descent II

Clever update scheme with low memory impact by storing :

- current **residual** in a variable $r^{(k)}$ (size n vector)
- current **coefficient** in $\theta^{(k)}$ (size p vector)

Coordinate descent for least square

Input: Observations y, features $X = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$, initial $\theta^{(0)}$
Result: Vector $\theta^{(K)}$
while *not converged* **do**

  Pick a coordinate j

  $r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$

  $\theta_j^{(k+1)} \leftarrow S_{\lambda/\|\mathbf{x}_j\|^2}\left(\mathbf{x}_j^\top r^{\text{int}}/\|\mathbf{x}_j\|^2\right)$

  $r^{(k+1)} = r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$

end

Rem: computational simplification $\|\mathbf{x}_j\|_2^2 = 1$ (NORMALIZE !)
Rem: the residual update can be done in place

Similarity with Forward-Backward

$$\text{Update rule for CD}: \quad \theta_j^{(k+1)} \leftarrow S_{\lambda/\|\mathbf{x}_j\|^2}\left(\mathbf{x}_j^\top r/\|\mathbf{x}_j\|^2\right)$$
$$\text{Update rule for FB}: \quad \theta^{(k+1)} \leftarrow S_{\lambda/L}\left(X^\top r/L\right)$$

where L is the Lipschitz constant of $X^\top X$, $L = \lambda_{\max}(X^\top X)$

<u>Rem</u>: The Forward-Backward update could be really useful if the following operation can be performed efficiently :

$$\begin{cases} \mathbb{R}^n & \to \mathbb{R}^p \\ r & \mapsto X^\top \cdot r \end{cases}$$

Common examples includes : FFT, wavelet transform, etc.
<u>Rem</u>: Note that the residual is usually a full vector ($\neq$ sparse)

Optimisation

Other alternatives to obtained a/the Lasso solution

- LARS Efron *et al.* (2004) for full Lasso path
- Forward-Backward (ISTA, FISTA, *cf.* Beck et Teboulle(2009))
- Conditional Gradient / Frank-Wolfe (Jaggi (2013)) useful for distributed datasets

Références I

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

- A. Chambolle and C. Dossal.
  How to make sure the iterates of fista converge.
  2014.

- B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani.
  Least angle regression.
  Ann. Statist., 32(2) :407–499, 2004.
  With discussion, and a rejoinder by the authors.

- J-B. Hiriart-Urruty and C. Lemaréchal.
  Convex analysis and minimization algorithms. I, volume 305.
  Springer-Verlag, Berlin, 1993.

- M. Jaggi.
  Revisiting {Frank-Wolfe} : Projection-free sparse convex optimization.
  In *ICML*, pages 427–435, 2013.

Références II

- Y. Nesterov.

  Introductory lectures on convex optimization, volume 87 of *Applied Optimization*.

  Kluwer Academic Publishers, Boston, MA, 2004.

- P. Tseng.

  Convergence of a block coordinate descent method for nondifferentiable minimization.

  J. Optim. Theory Appl., 109(3) :475–494, 2001.