STAT 593
Robust statistics: Majorization Minimization

Joseph Salmon
http://josephsalmon.eu

Télécom Paristech, Institut Mines-Télécom
&
University of Washington, Department of Statistics
(Visiting Assistant Professor)

Outline

Smooth function: gradient Lipschitz

Quadratic majorization

If f is convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|$$

then the following holds: $\forall(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$0 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq \frac{L}{2} \left\|\theta' - \theta\right\|^2$$

Smooth function: gradient Lipschitz

Quadratic majorization

If f is convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\|$$

then the following holds: $\quad \forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d,$

$$0 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq \frac{L}{2} \left\| \theta' - \theta \right\|^2$$

Rem: positivity is a consequence of convexity. The second inequality will be proved later on.

Smooth function: gradient Lipschitz

Quadratic majorization

If f is convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\|$$

then the following holds: $\quad \forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d,$

$$0 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq \frac{L}{2} \left\|\theta' - \theta\right\|^2$$

Rem: positivity is a consequence of convexity. The second inequality will be proved later on.

Rem: if f is twice differentiable $\nabla^2 f \preceq L \cdot \mathrm{Id}_d$ in the sense that $L \cdot \mathrm{Id}_d - \nabla^2 f$ is semi-definite positive, then ∇f is L-Lipschitz

Majorization/minimization

Fix θ^0, and assume the previous inequality holds for any $\theta \in \mathbb{R}^d$:

$$f(\theta) - f(\theta^0) - \langle \nabla f(\theta^0), \theta - \theta^0 \rangle \leq \frac{L}{2} \left\| \theta^0 - \theta \right\|^2$$

yields

$$\begin{aligned} f(\theta) \leq & f(\theta^0) + \langle \nabla f(\theta^0), \theta - \theta^0 \rangle + \frac{L}{2} \| \theta^0 - \theta \|^2 \\ = & \tfrac{L}{2} \left\| \theta^0 - \tfrac{1}{L} \nabla f(\theta^0) - \theta \right\|^2 + f(\theta^0) - \tfrac{1}{2L} \left\| \nabla f(\theta^0) \right\|^2 := Q_L(\theta^0, \theta) \end{aligned}$$

Hence : $\quad \forall \theta \in \mathbb{R}^d, \quad \begin{cases} Q_L(\theta^0, \theta^0) = f(\theta^0) \\ f(\theta) \leq Q_L(\theta^0, \theta) \end{cases}$. This leads to a tight upper bound that can be simply minimized, since

$$\arg\min_{\theta \in \mathbb{R}^d} Q_L(\theta^0, \theta) = \theta^0 - \frac{1}{L} \nabla f(\theta^0)$$

Example on a simple case: Huber function

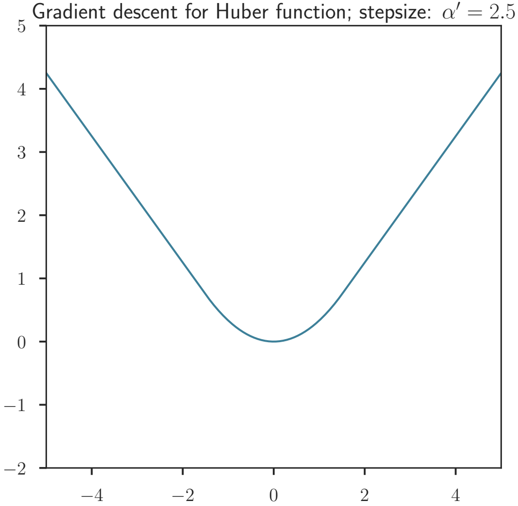
Remind that

$$\rho_\alpha = \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \le \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

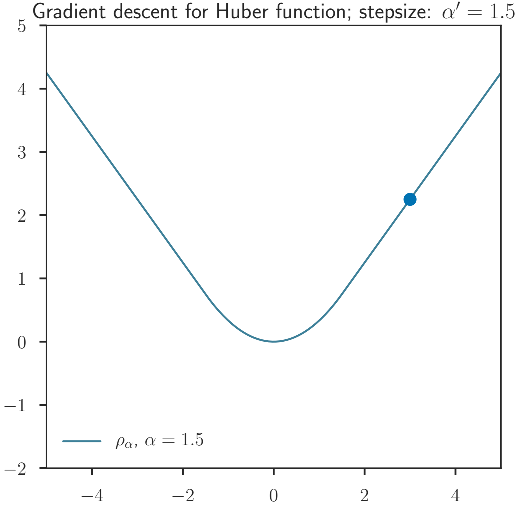
Then, one can show[1] that this is a convex function with gradient L-Lipschitz for $L = \frac{1}{\alpha}$.

[1] A. Beck and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.

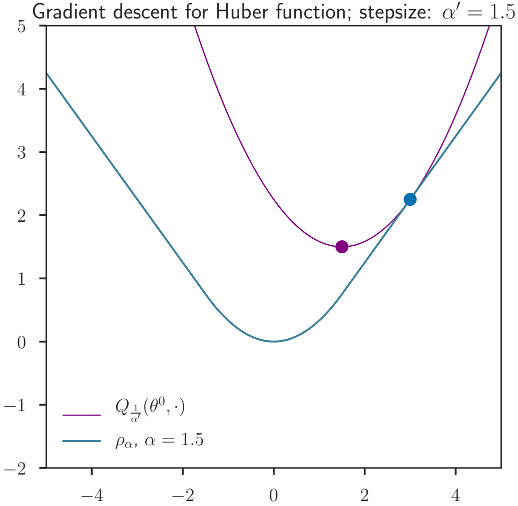
Majorization / Minimization



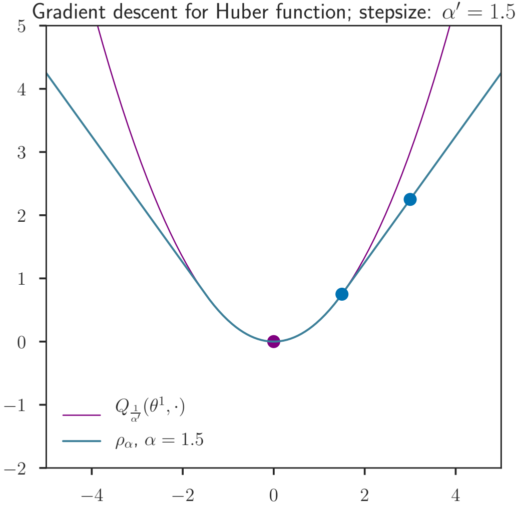
Majorization / Minimization



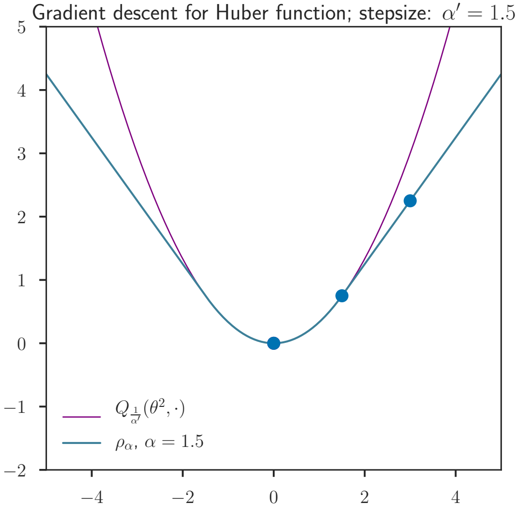
Majorization / Minimization



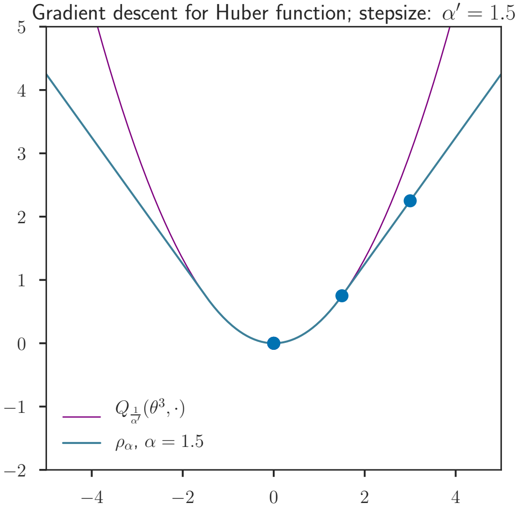
Majorization / Minimization



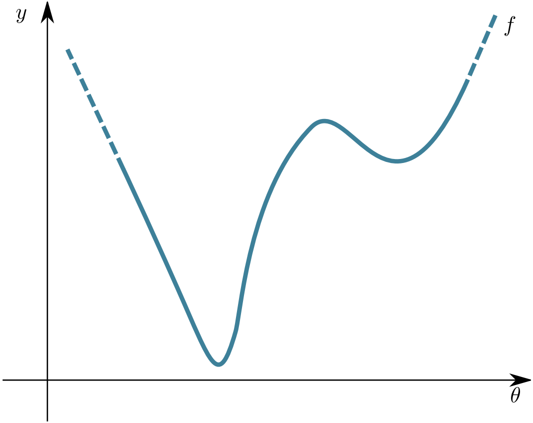
Majorization / Minimization



Majorization / Minimization

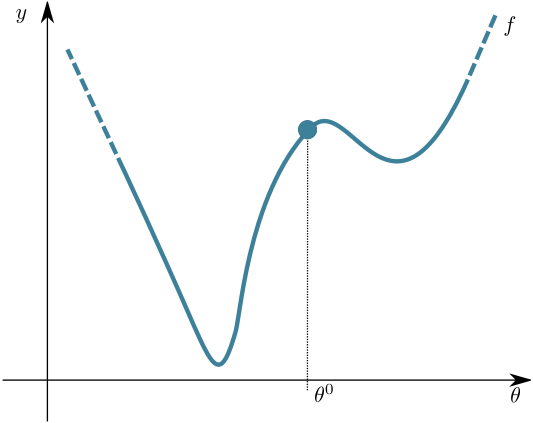


Majorization / Minimization: visually



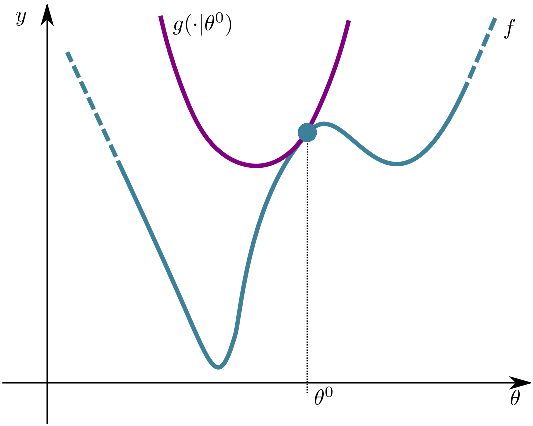
Original function

Majorization / Minimization: visually



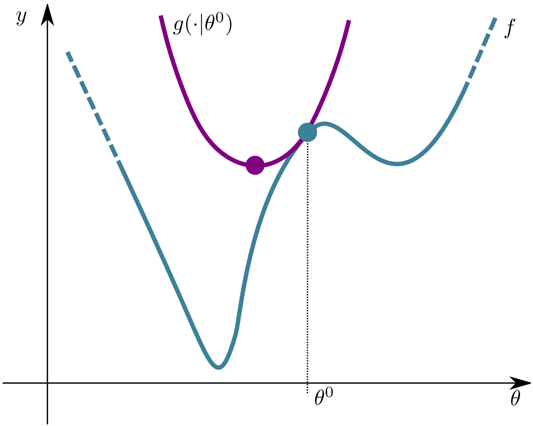
Initialize

Majorization / Minimization: visually



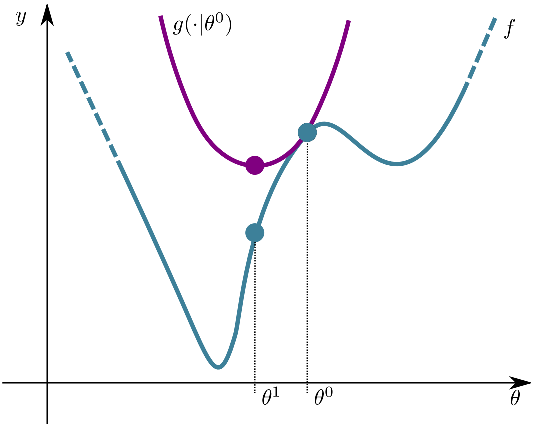
Majorize

Majorization / Minimization: visually



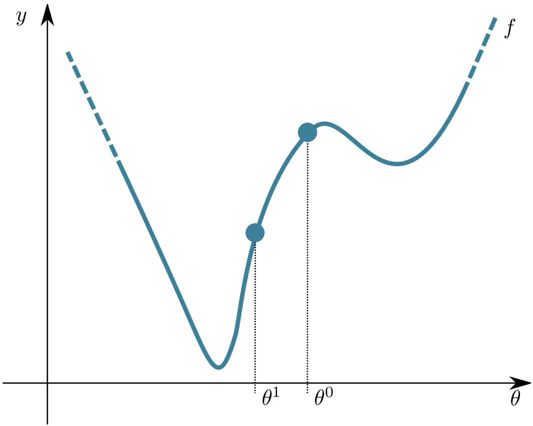
Minimize

Majorization / Minimization: visually



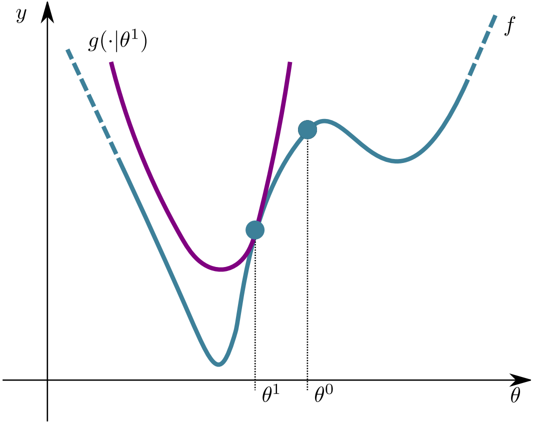
Update

Majorization / Minimization: visually



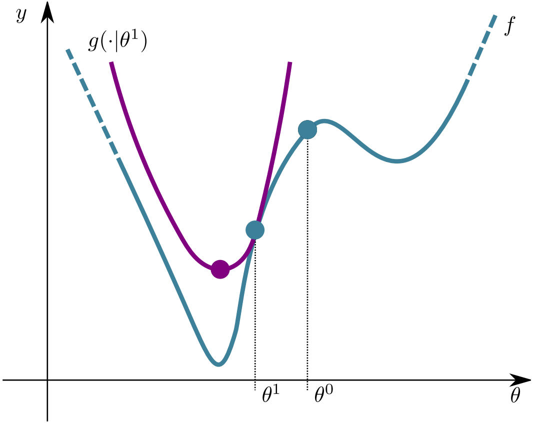
Update

Majorization / Minimization: visually



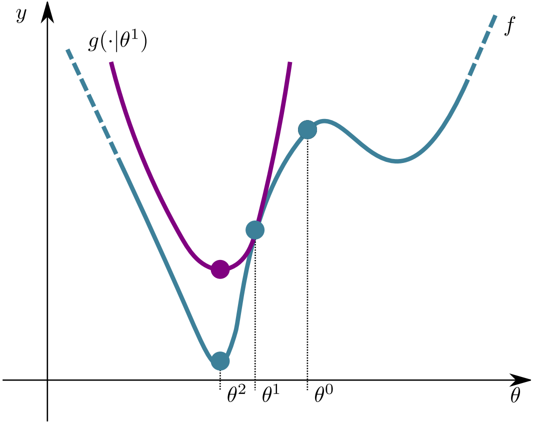
Majorize

Majorization / Minimization: visually



Minimize

Majorization / Minimization: visually



Update

Majorization / Minimization: formally

Objective: find a minimizer of a function f

Tool: at each point θ^t proceed as follows:

- Provide a "majorization" function $\theta \to g(\theta|\theta^t)$ satisfying:

$$\begin{cases} f(\theta) \leq g(\theta|\theta^t), \forall\theta & : \quad \text{domination / upper bound} \\ f(\theta^t) = g(\theta^t|\theta^t) & : \quad \text{tangency / tightness at } \theta^t \end{cases}$$

- Minimize the upper bound and obtain

$$\theta^{t+1} \in \arg\min_{\theta \in \mathbb{R}^d} g(\theta|\theta^t)$$

Rem: we say that $g(\cdot|\theta^t)$ is a surrogate of f at θ^t

Majorization / Minimization: Algorithm

Algorithm: MAXIMIZATION MINIMIZATION

---

input : max. iterations $t_{\max}$, stopping criterion ε

init : θ^0

for $1 \leq t \leq t_{\max}$ **do**

  Break if stopping criterion smaller than ε

  Find a majorization function: $g(\cdot|\theta^t)$

  Minimize this bound: $\theta^{t+1} \leftarrow \underset{\theta \in \mathbb{R}^d}{\arg\min}\, g(\theta|\theta^t)$

return $\theta^{t_{\max}}$ *"close" to a local minimum of f*

Convergence property[2]

Theorem

The maximization/minimization algorithm is a descent method:

$$\forall t \geq 1, \quad f(\theta^{t+1}) \leq f(\theta^{t})$$

Hence, provided that f is lower bounded the algorithm converges.

[2] K. Lange. *MM optimization algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016, pp. ix+223.

Convergence property[2]

Theorem

The maximization/minimization algorithm is a descent method:

$$\forall t \geq 1, \quad f(\theta^{t+1}) \leq f(\theta^t)$$

Hence, provided that f is lower bounded the algorithm converges.

Proof:

[2] K. Lange. *MM optimization algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016, pp. ix+223.

Convergence property[2]

Theorem

The maximization/minimization algorithm is a descent method:

$$\forall t \geq 1, \quad f(\theta^{t+1}) \leq f(\theta^t)$$

Hence, provided that f is lower bounded the algorithm converges.

Proof:

$$f(\theta^{t+1}) \leq g(\theta^{t+1}|\theta^t) \qquad \text{(Majorization at } \theta^t\text{)}$$

[2] K. Lange. *MM optimization algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016, pp. ix+223.

Convergence property[2]

Theorem

The maximization/minimization algorithm is a descent method:

$$\forall t \geq 1, \quad f(\theta^{t+1}) \leq f(\theta^t)$$

Hence, provided that f is lower bounded the algorithm converges.

Proof:

$$\begin{aligned} f(\theta^{t+1}) &\leq g(\theta^{t+1}|\theta^t) && \text{(Majorization at } \theta^t\text{)} \\ &\leq g(\theta^t|\theta^t) && \text{(Minimization definition of } \theta^{t+1}\text{)} \end{aligned}$$

[2] K. Lange. *MM optimization algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016, pp. ix+223.

Convergence property[2]

Theorem

The maximization/minimization algorithm is a descent method:

$$\forall t \geq 1, \quad f(\theta^{t+1}) \leq f(\theta^t)$$

Hence, provided that f is lower bounded the algorithm converges.

<u>Proof</u>:

$$\begin{aligned} f(\theta^{t+1}) &\leq g(\theta^{t+1}|\theta^t) && \text{(Majorization at } \theta^t\text{)} \\ &\leq g(\theta^t|\theta^t) && \text{(Minimization definition of } \theta^{t+1}\text{)} \\ &= f(\theta^t) && \text{(tightness at } \theta^t\text{)} \end{aligned}$$

[2] K. Lange. *MM optimization algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016, pp. ix+223.

Various examples: gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta)}$

Properties: f is convex with gradient L-Lipschitz

Various examples: gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta)}$

Properties: f is convex with gradient L-Lipschitz

Surrogate:

$$g(\theta|\theta^t) = f(\theta^t) + \langle \nabla f(\theta^t), \theta - \theta^t \rangle + \frac{L}{2} \|\theta^t - \theta\|^2$$

Various examples: gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta)}$

Properties: f is convex with gradient L-Lipschitz

Surrogate:

$$g(\theta|\theta^t) = f(\theta^t) + \langle \nabla f(\theta^t), \theta - \theta^t \rangle + \frac{L}{2} \|\theta^t - \theta\|^2$$

Update rule : $\boxed{\theta^{t+1} = \theta^t - \frac{1}{L} \nabla f(\theta^t)}$

Various examples: proximal gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) + \psi(\theta)}$

Properties: f convex, gradient L-Lipschitz; ψ convex s.t. prox_ψ (the **proximal** operator[3] of ψ) has a closed-form, where

$$\operatorname{prox}_\psi\left(\theta^0\right) = \arg\min_{\theta \in \mathbb{R}^d} \tfrac{1}{2}\|\theta - \theta^0\|^2 + \psi(\theta)$$

[3] J.-J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien". In: *C. R. Acad. Sci. Paris* 255 (1962), pp. 2897–2899.

Various examples: proximal gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) + \psi(\theta)}$

Properties: f convex, gradient L-Lipschitz; ψ convex s.t. prox_ψ (the **proximal** operator[3] of ψ) has a closed-form, where

$$\mathrm{prox}_\psi\left(\theta^0\right) = \arg\min_{\theta \in \mathbb{R}^d} \tfrac{1}{2}\|\theta - \theta^0\|^2 + \psi(\theta)$$

Surrogate: $g(\theta|\theta^t) = f(\theta^t) + \langle \nabla f(\theta^t), \theta - \theta^t \rangle + \frac{L\|\theta^t - \theta\|^2}{2} + \psi(\theta)$

[3] J.-J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien". In: *C. R. Acad. Sci. Paris* 255 (1962), pp. 2897–2899.

Various examples: proximal gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) + \psi(\theta)}$

Properties: f convex, gradient L-Lipschitz; ψ convex s.t. prox_ψ (the **proximal** operator[3] of ψ) has a closed-form, where

$$\mathrm{prox}_\psi\left(\theta^0\right) = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|\theta - \theta^0\|^2 + \psi(\theta)$$

Surrogate: $g(\theta|\theta^t) = f(\theta^t) + \langle \nabla f(\theta^t), \theta - \theta^t \rangle + \frac{L\|\theta^t - \theta\|^2}{2} + \psi(\theta)$

Update rule : $\boxed{\theta^{t+1} = \mathrm{prox}_{\frac{\psi}{L}}\left(\theta^t - \frac{1}{L}\nabla f(\theta^t)\right)}$

[3] J.-J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien". In: *C. R. Acad. Sci. Paris* 255 (1962), pp. 2897–2899.

Various examples: proximal gradient descent

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) + \psi(\theta)}$

Properties: f convex, gradient L-Lipschitz; ψ convex s.t. prox_ψ (the **proximal** operator(3) of ψ) has a closed-form, where

$$\mathrm{prox}_\psi\left(\theta^0\right) = \arg\min_{\theta \in \mathbb{R}^d} \tfrac{1}{2}\|\theta - \theta^0\|^2 + \psi(\theta)$$

Surrogate: $g(\theta|\theta^t) = f(\theta^t) + \langle \nabla f(\theta^t), \theta - \theta^t \rangle + \frac{L\|\theta^t - \theta\|^2}{2} + \psi(\theta)$

Update rule : $\boxed{\theta^{t+1} = \mathrm{prox}_{\frac{\psi}{L}}\left(\theta^t - \tfrac{1}{L}\nabla f(\theta^t)\right)}$

Proof (*cf.* gradient descent): $\theta^{t+1} = \arg\min_{\theta \in \mathbb{R}^d} \frac{L\|\theta^t - \frac{1}{L}\nabla f(\theta^t) - \theta\|^2}{2} + \psi(\theta)$

(3) J.-J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien". In: *C. R. Acad. Sci. Paris* 255 (1962), pp. 2897–2899.

More on proximal methods

This is particularly popular for solving Lasso type problems (in image/signal processing, in statistics / ML coordinate descent more popular):

$$\hat{\theta} \in \arg\min_{\theta} \left(\frac{1}{2} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right)$$

(4) N. Parikh et al. "Proximal algorithms". In: *Foundations and Trends in Machine Learning* 1.3 (2013), pp. 1–108.

More on proximal methods

This is particularly popular for solving Lasso type problems (in image/signal processing, in statistics / ML coordinate descent more popular):

$$\hat{\theta} \in \arg\min_{\theta} \left(\frac{1}{2} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right)$$

List of proximal operators[(4)]:

- projection over a closed convex set

(4) N. Parikh et al. "Proximal algorithms". In: *Foundations and Trends in Machine Learning* 1.3 (2013), pp. 1–108.

More on proximal methods

This is particularly popular for solving Lasso type problems (in image/signal processing, in statistics / ML coordinate descent more popular):

$$\hat{\theta} \in \arg\min_{\theta} \left(\frac{1}{2} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right)$$

List of proximal operators[(4)]:

- projection over a closed convex set
- (block) soft-thresholding operator

[(4)] N. Parikh et al. "Proximal algorithms". In: *Foundations and Trends in Machine Learning* 1.3 (2013), pp. 1–108.

More on proximal methods

This is particularly popular for solving Lasso type problems (in image/signal processing, in statistics / ML coordinate descent more popular):

$$\hat{\theta} \in \arg\min_{\theta} \left(\frac{1}{2} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right)$$

List of proximal operators[4]:

- projection over a closed convex set
- (block) soft-thresholding operator
- shrinkage operator (Ridge)

[4] N. Parikh et al. "Proximal algorithms". In: *Foundations and Trends in Machine Learning* 1.3 (2013), pp. 1–108.

More on proximal methods

This is particularly popular for solving Lasso type problems (in image/signal processing, in statistics / ML coordinate descent more popular):

$$\hat{\theta} \in \arg\min_{\theta} \left(\frac{1}{2} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right)$$

List of proximal operators[4]:

- projection over a closed convex set
- (block) soft-thresholding operator
- shrinkage operator (Ridge)
- etc.

[4] N. Parikh et al. "Proximal algorithms". In: *Foundations and Trends in Machine Learning* 1.3 (2013), pp. 1–108.

Various examples: coordinate descent

The variant of the Gradient Descent and Proximal apply coordinate-wise, can be encompassed in a similar way[5]

[5] Z. Peng et al. "Coordinate-friendly structures, algorithms and applications". In: *Ann. Math. Sci. Appl.* 1.1 (2016), pp. 57–119. ISSN: 2380-288X; 2380-2898/e.

Various examples: Difference of convex (DC-Programming)

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) - h(\theta)}$

Properties: f and h are convex and ∇h exists

(6) H. Zou. "The adaptive lasso and its oracle properties". In: *J. Amer. Statist. Assoc.* 101.476 (2006), pp. 1418–1429.

(7) E. J. Candès, M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In: *J. Fourier Anal. Applicat.* 14.5-6 (2008), pp. 877–905.

Various examples: Difference of convex (DC-Programming)

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) - h(\theta)}$

Properties: f and h are convex and ∇h exists

Surrogate:

$$g(\theta|\theta^t) = f(\theta) - h(\theta^t) - \langle \nabla h(\theta^t), \theta - \theta^t \rangle$$

[6] H. Zou. "The adaptive lasso and its oracle properties". In: *J. Amer. Statist. Assoc.* 101.476 (2006), pp. 1418–1429.

[7] E. J. Candès, M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In: *J. Fourier Anal. Applicat.* 14.5-6 (2008), pp. 877–905.

Various examples: Difference of convex (DC-Programming)

Optimization problem: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) - h(\theta)}$

Properties: f and h are convex and ∇h exists

Surrogate:

$$g(\theta|\theta^t) = f(\theta) - h(\theta^t) - \langle \nabla h(\theta^t), \theta - \theta^t \rangle$$

Usage: adaptive Lasso[6] / re-weighted[7] ℓ_1

(6) H. Zou. "The adaptive lasso and its oracle properties". In: *J. Amer. Statist. Assoc.* 101.476 (2006), pp. 1418–1429.

(7) E. J. Candès, M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In: *J. Fourier Anal. Applicat.* 14.5-6 (2008), pp. 877–905.

Geometric median / Weiszfeld algorithm[8]

Definition

(Geometric) Median : $$\mathrm{Med}_n(\mathbf{x}) \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \|\theta - x_i\| = f(\theta)$$

With concavity of $\sqrt{\cdot}$ over $\mathbb{R}_+$, on has:

$$\forall x \geq 0, y > 0, \quad \sqrt{x} \leq \sqrt{y} + \frac{1}{2\sqrt{y}}(x - y)$$

leading to the following majorization function for f:

$$g(\theta|\theta^t) = \sum_{i=1}^{n} \left(\|\theta^t - x_i\| + \frac{\|\theta - x_i\|^2 - \|\theta^t - x_i\|^2}{2\,\|\theta^t - x_i\|} \right)$$

[8] E. Weiszfeld. "Sur le point pour lequel la somme des distances de n points donnés est minimum". In: *Tohoku Mathematical Journal, First Series* 43 (1937), pp. 355–386.

Geometric median / Weiszfeld algorithm (bis)

$$
\begin{aligned}
\theta^{t+1} &= \underset{\theta \in \mathbb{R}^d}{\arg\min}\, g(\theta|\theta^t) \\
&= \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \left(\|\theta^t - x_i\| + \frac{\|\theta - x_i\|^2 - \|\theta^t - x_i\|^2}{2\,\|\theta^t - x_i\|} \right)
\end{aligned}
$$

Geometric median / Weiszfeld algorithm (bis)

$$
\begin{aligned}
\theta^{t+1} &= \arg\min_{\theta \in \mathbb{R}^d} g(\theta|\theta^t) \\
&= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \left(\|\theta^t - x_i\| + \frac{\|\theta - x_i\|^2 - \|\theta^t - x_i\|^2}{2\,\|\theta^t - x_i\|} \right) \\
&= \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^n \frac{\|\theta - x_i\|^2}{\|\theta^t - x_i\|}
\end{aligned}
$$

Geometric median / Weiszfeld algorithm (bis)

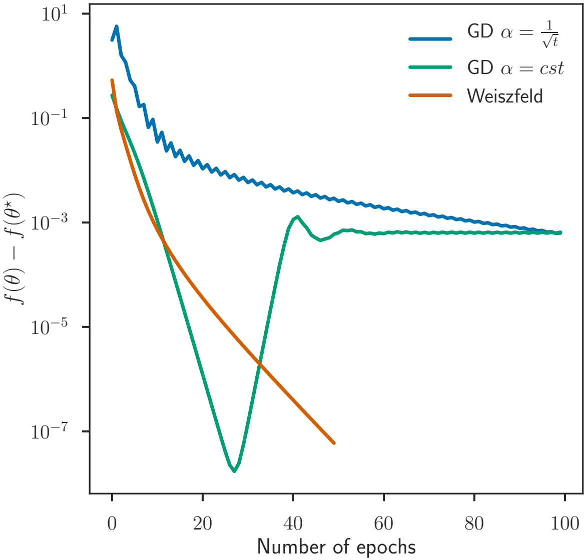
$$
\begin{aligned}
\theta^{t+1} &= \arg\min_{\theta \in \mathbb{R}^d} g(\theta | \theta^t) \\
&= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\|\theta^t - x_i\| + \frac{\|\theta - x_i\|^2 - \|\theta^t - x_i\|^2}{2\,\|\theta^t - x_i\|} \right) \\
&= \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \frac{\|\theta - x_i\|^2}{\|\theta^t - x_i\|} \\
&= \sum_{i=1}^{n} \frac{w_i^t}{\sum_{i'=1}^{n} w_{i'}^t} x_i, \quad \text{where} \quad w_i^t = \frac{1}{\|\theta^t - x_i\|}
\end{aligned}
$$

Geometric median / Weiszfeld algorithm (bis)

$$
\begin{aligned}
\theta^{t+1} &= \underset{\theta \in \mathbb{R}^d}{\arg\min}\, g(\theta|\theta^t) \\
&= \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \left(\|\theta^t - x_i\| + \frac{\|\theta - x_i\|^2 - \|\theta^t - x_i\|^2}{2\,\|\theta^t - x_i\|} \right) \\
&= \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \frac{1}{2} \sum_{i=1}^{n} \frac{\|\theta - x_i\|^2}{\|\theta^t - x_i\|} \\
&= \sum_{i=1}^{n} \frac{w_i^t}{\sum_{i'=1}^{n} w_{i'}^t} x_i, \quad \text{where} \quad w_i^t = \frac{1}{\|\theta^t - x_i\|}
\end{aligned}
$$

Rem: to avoid any problem at 0, substitute $w_i^t = \sqrt{\|\theta^t - x_i\|^2 + \epsilon}$

Comparisons



Other non-convex M-estimators

M-estimator associated to a function ρ:

$$\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \rho(x_i - \theta)$$

and assume one can write $\rho(x) = f(\|x\|^2)$ with f concave. Let us write $W(x) = f'(\|x\|^2)$.

Surrogate:

$$g(\theta|\theta^t) = \sum_{i=1}^n \left(\rho(x_i - \theta^t) + W(x_i - \theta^t)\left[\|x_i - \theta\|^2 - \|x_i - \theta^t\|^2\right]\right)$$

Update rule: $\theta^{t+1} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n W(x_i - \theta^t)\|x_i - \theta\|^2$

Proof

Start using the concavity of f :

$$f(s) \leq f(s^0) + (s - s^0) f'(s^0), \forall s, s^0$$

Proof

Start using the concavity of f :

$$f(s) \leq f(s^0) + (s - s^0) f'(s^0), \forall s, s^0$$

Apply the former for $s = \|x_i - \theta\|^2$ and $s^0 = (\|x_i - \theta^t\|^2)$ yields:

$$\begin{aligned}
\rho(x_i - \theta) =& f(\|x_i - \theta\|^2) \\
\leq& f(\|x_i - \theta^t\|^2) + f'(\|x_i - \theta^t\|^2) \left[\|x_i - \theta\|^2 - \|x_i - \theta^t\|^2 \right]
\end{aligned}$$

Proof

Start using the concavity of f :

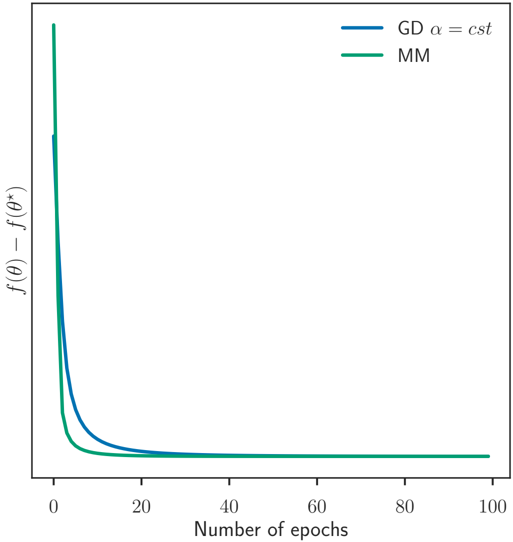
$$f(s) \leq f(s^0) + (s - s^0) f'(s^0), \forall s, s^0$$

Apply the former for $s = \|x_i - \theta\|^2$ and $s^0 = (\|x_i - \theta^t\|^2)$ yields:

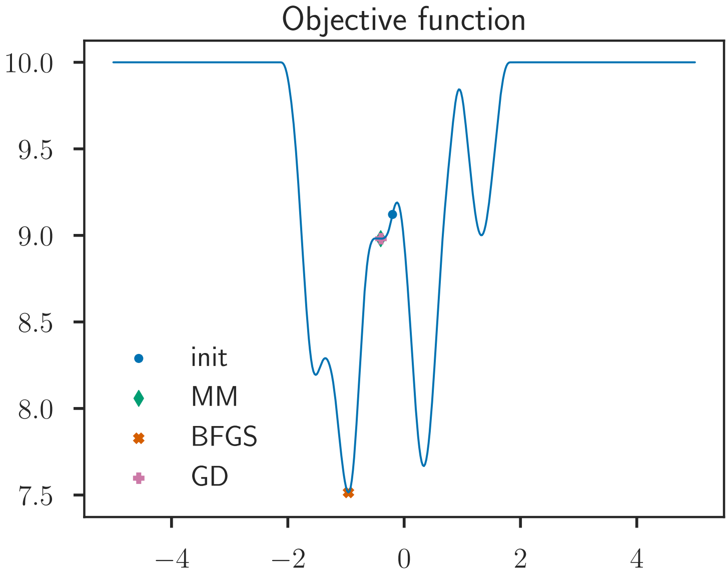
$$\begin{aligned}
\rho(x_i - \theta) =& f(\|x_i - \theta\|^2) \\
\leq& f(\|x_i - \theta^t\|^2) + f'(\|x_i - \theta^t\|^2) \left[\|x_i - \theta\|^2 - \|x_i - \theta^t\|^2\right] \\
=& \rho(x_i - \theta^t) + W(x_i - \theta^t) \left[\|x_i - \theta\|^2 - \|x_i - \theta^t\|^2\right]
\end{aligned}$$

where we have used $W(x) = f'(\|x\|^2)$ for the last equality

Comparisons



Comparisons



More references on the field

- (proximal) gradient descent and MM: Beck and Teboulle (2009)
- Concomitant MM approaches: Wolke and Schwetlick (1988)

References I

- Beck, A. and M. Teboulle. "Gradient-based algorithms with applications to signal-recovery problems". In: *Convex Optimization in Signal Processing and Communications*. Ed. by D. P. Palomar and Y. C. Eldar. Cambridge University Press, 2009, pp. 42–88.
- – ."Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.
- Candès, E. J., M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In: *J. Fourier Anal. Applicat.* 14.5-6 (2008), pp. 877–905.
- Lange, K. *MM optimization algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016, pp. ix+223.
- Moreau, J.-J. "Fonctions convexes duales et points proximaux dans un espace hilbertien". In: *C. R. Acad. Sci. Paris* 255 (1962), pp. 2897–2899.

References II

- Parikh, N. et al. "Proximal algorithms". In: *Foundations and Trends in Machine Learning* 1.3 (2013), pp. 1–108.
- Peng, Z. et al. "Coordinate-friendly structures, algorithms and applications". In: *Ann. Math. Sci. Appl.* 1.1 (2016), pp. 57–119. ISSN: 2380-288X; 2380-2898/e.
- Weiszfeld, E. "Sur le point pour lequel la somme des distances de n points donnés est minimum". In: *Tohoku Mathematical Journal, First Series* 43 (1937), pp. 355–386.
- Wolke, R. and H. Schwetlick. "Iteratively Reweighted Least Squares: Algorithms, Convergence Analysis, and Numerical Comparisons". In: *SIAM Journal on Scientific and Statistical Computing* 9.5 (1988), pp. 907–921.
- Zou, H. "The adaptive lasso and its oracle properties". In: *J. Amer. Statist. Assoc.* 101.476 (2006), pp. 1418–1429.