Proximal Neural Networks: Wedding Variational Methods and Artificial Intelligence

II – Proximal algorithms

Audrey Repetti[†] , Nelly Pustelnik[◇], Jean-Christophe Pesquet[*]

[*] CentraleSupélec, Université Paris-Saclay, Inria, Gif-sur-Yvettes, France
[◇] CNRS, ENS Lyon, Lyon, France
[†] Heriot-Watt University & Maxwell Institute for Mathematical Sciences, Edinburgh, UK

Tutorial – Eusipco 2025 – Palermo, Italy

Unified framework

Inference framework: feed-forward NN

$$(\forall \boldsymbol{x}^{[0]} \in \mathbb{R}^{N_0}) \qquad \boldsymbol{x}^{[K]} = \mathfrak{L}_{\Theta}^{K}(\boldsymbol{x}^{[0]}) = \mathfrak{T}_{\Theta_K} \circ \ldots \circ \mathfrak{T}_{\Theta_1}(\boldsymbol{x}^{[0]}),$$

Layer/iteration

$$\mathfrak{T}_{\Theta_k} \colon \mathbb{R}^{N_{k-1}} \to \mathbb{R}^{N_k} \colon \boldsymbol{x} \mapsto \mathfrak{D}_{\Lambda_k}(\mathbf{L}_k \boldsymbol{x} + \boldsymbol{b}_k),$$

- $\mathbf{L}_k \colon \mathbb{R}^{N_{k-1}} \to \mathbb{R}^{N_k}$: linear operator,
- $\boldsymbol{b}_k \in \mathbb{R}^{N_k}$: shift parameter,
- $\mathfrak{D}_{\Lambda_k} \colon \mathbb{R}^{N_k} \to \mathbb{R}^{N_k}$: nonlinear operator parametrized by Λ_k.

Parameters: $\Theta = \cup_{k=1}^{K} \Theta_k$ with $\Theta_k = \{\Lambda_k, \mathbf{L}_k, \boldsymbol{b}_k\}$.

Basic convex analysis tools

- **Hilbert space** $\mathcal{H}$
- **Moreau subdifferential**
  Let $f\colon \mathcal{H} \to (-\infty, +\infty]$ and $\boldsymbol{x} \in \mathcal{H}$

$$\boxed{\partial f(\boldsymbol{x}) = \{\boldsymbol{t} \in \mathcal{H} \mid (\forall \boldsymbol{y} \in \mathcal{H}) \;\; f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{t} \mid \boldsymbol{y} - \boldsymbol{x} \rangle\}.}$$

- $\Gamma_0(\mathcal{H})$: class of lower-semicontinuous convex functions, finite at least at one point (proper)
- If $f \in \Gamma_0(\mathcal{H})$ is Gâteaux-differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$
- Fermat's rule: $\boxed{\widehat{\boldsymbol{x}} \in \operatorname{Argmin} \; f \quad \Leftrightarrow \quad 0 \in \partial f(\widehat{\boldsymbol{x}})}$

Basic convex analysis tools

- **Hilbert space** $\mathcal{H}$
- **Proximity operator**
  Let $f\colon \mathcal{H} \to (-\infty, +\infty]$ and $\boldsymbol{x} \in \mathcal{H}$

$$\boxed{\operatorname{prox}_f(\boldsymbol{x}) \in \operatorname*{Argmin}_{\boldsymbol{y}\in\mathcal{H}} \; \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + f(\boldsymbol{y}).}$$

- $\boldsymbol{p} = \operatorname{prox}_f(\boldsymbol{x}) \quad \Leftrightarrow \quad \boldsymbol{x}-\boldsymbol{p} \in \partial f(\boldsymbol{p}) \quad \Leftrightarrow \quad \boldsymbol{p} \in (\mathrm{Id} + \partial f)^{-1}(\boldsymbol{x})$
- See https://proximity-operator.net for the expression/code of prox_f for many functions f

Basic convex analysis tools

- **Hilbert space $\mathcal{H}$**
- **Proximity operator**
  Let $f\colon \mathcal{H} \to (-\infty, +\infty]$ and $\boldsymbol{x} \in \mathcal{H}$

$$\boxed{\operatorname{prox}_f(\boldsymbol{x}) \in \underset{\boldsymbol{y}\in\mathcal{H}}{\operatorname{Argmin}} \ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + f(\boldsymbol{y}).}$$

- f is **ρ-convex** if $f - \frac{\rho}{2}\|\cdot\|^2$ is convex
  if $\rho > 0$, f is **ρ-strongly convex**
  if $\rho < 0$, f is **$(-\rho)$-weakly convex**
- If f is proper lower-semicontinuous and ρ-convex with $\rho > -1$, then $\operatorname{prox}_f(\boldsymbol{x})$ is uniquely defined for every $\boldsymbol{x} \in \mathcal{H}$.

Basic convex analysis tools

- **Hilbert space** $\mathcal{H}$

- **Proximity operator**
  Let $f\colon \mathcal{H} \to (-\infty, +\infty]$ and $\boldsymbol{x} \in \mathcal{H}$

$$\boxed{\operatorname{prox}_f(\boldsymbol{x}) \in \underset{\boldsymbol{y}\in\mathcal{H}}{\operatorname{Argmin}} \ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + f(\boldsymbol{y}).}$$

- **Moreau envelope**

$$(\forall \boldsymbol{x} \in \mathcal{H}) \quad \widetilde{f}(\boldsymbol{x}) = \inf_{\boldsymbol{y}\in\mathcal{H}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + f(\boldsymbol{y}).$$

- $\operatorname{Argmin} \ f = \operatorname{Argmin} \ \widetilde{f}$

- If f is proper lower-semicontinuous and ρ-convex with $\rho > -1$, then $\widetilde{f}$ is $\rho/(1+\rho)$-convex with Lipschitz continuous gradient $\nabla \widetilde{f} = \operatorname{Id} - \operatorname{prox}_f$.

Fixed point algorithm: zeros and fixed points

Let $\mathcal{H}$ be a Hilbert space. Let $\boldsymbol{\Phi}\colon \mathcal{H} \to 2^{\mathcal{H}}$ and $\mathfrak{T}\colon \mathcal{H} \to 2^{\mathcal{H}}$.
The set of **fixed points** of $\mathfrak{T}$ is : $\operatorname{Fix} \mathfrak{T} = \{\boldsymbol{x} \in \mathcal{H} \,|\, \boldsymbol{x} \in \mathfrak{T}\boldsymbol{x}\}$.
The set of **zeros** of $\boldsymbol{\Phi}$ is : zer $\boldsymbol{\Phi} = \{\boldsymbol{x} \in \mathcal{H} \,|\, 0 \in \boldsymbol{\Phi}\boldsymbol{x}\}$.

Goal: Find $\boldsymbol{\Phi}$ and $\mathfrak{T}$ such that $\operatorname{Argmin}\ f = \operatorname{zer} \boldsymbol{\Phi} = \operatorname{Fix} \mathfrak{T}$

Fixed point algorithm: zeros and fixed points

Let $\mathcal{H}$ be a Hilbert space. Let $\boldsymbol{\Phi}\colon \mathcal{H} \to 2^{\mathcal{H}}$ and $\mathfrak{T}\colon \mathcal{H} \to 2^{\mathcal{H}}$.
The set of **fixed points** of $\mathfrak{T}$ is : $\mathrm{Fix}\, \mathfrak{T} = \{\boldsymbol{x} \in \mathcal{H} \,|\, \boldsymbol{x} \in \mathfrak{T}\boldsymbol{x}\}$.
The set of **zeros** of $\boldsymbol{\Phi}$ is : zer $\boldsymbol{\Phi} = \{\boldsymbol{x} \in \mathcal{H} \,|\, 0 \in \boldsymbol{\Phi}\boldsymbol{x}\}$.

Goal: Find $\boldsymbol{\Phi}$ and $\mathfrak{T}$ such that $\mathrm{Argmin}\ f = \mathrm{zer}\,\boldsymbol{\Phi} = \mathrm{Fix}\,\mathfrak{T}$

Example 1: gradient descent
If f differentiable and convex,

$$\boldsymbol{\Phi} = \nabla f, \quad \mathfrak{T} = \mathrm{Id} - \nabla f$$

Fixed point algorithm: zeros and fixed points

Let $\mathcal{H}$ be a Hilbert space. Let $\mathbf{\Phi}\colon \mathcal{H} \to 2^{\mathcal{H}}$ and $\mathfrak{T}\colon \mathcal{H} \to 2^{\mathcal{H}}$.
The set of **fixed points** of $\mathfrak{T}$ is : $\mathrm{Fix}\,\mathfrak{T} = \{\boldsymbol{x} \in \mathcal{H} \,|\, \boldsymbol{x} \in \mathfrak{T}\boldsymbol{x}\}$.
The set of **zeros** of $\mathbf{\Phi}$ is : zer $\mathbf{\Phi} = \{\boldsymbol{x} \in \mathcal{H} \,|\, 0 \in \mathbf{\Phi}\boldsymbol{x}\}$.

Goal: Find $\mathbf{\Phi}$ and $\mathfrak{T}$ such that $\mathrm{Argmin}\ f = \mathrm{zer}\ \mathbf{\Phi} = \mathrm{Fix}\ \mathfrak{T}$

Example 2: proximal point

$$\widehat{\boldsymbol{x}} \in \mathrm{Argmin}\ f \quad \Leftrightarrow \quad 0 \in \partial f(\widehat{x}) \quad \Leftrightarrow \quad \widehat{x} - \widehat{x} \in \partial f(\widehat{x}) \quad \Leftrightarrow \quad \widehat{x} \in \mathrm{prox}_f(\widehat{x})$$

$$\Rightarrow \mathbf{\Phi} = \partial f, \quad \mathfrak{T} = \mathrm{prox}_f = \mathrm{Id} - \nabla \widetilde{f}$$

Question: How to find a minimizer $\widehat{x}$?

Fixed point algorithm: convergence

Let $\mathcal{H}$ be a Hilbert space, $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ and $\widehat{\boldsymbol{x}} \in \mathcal{H}$.

- $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ **converges strongly** to $\widehat{\boldsymbol{x}}$ if

$$\lim_{k\to\infty} \|\boldsymbol{x}^{[k]} - \widehat{\boldsymbol{x}}\| = 0.$$

It is denoted by $\boldsymbol{x}^{[k]} \to \widehat{\boldsymbol{x}}$.

- $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ **converges weakly** to $\widehat{\boldsymbol{x}}$ if

$$(\forall \boldsymbol{u} \in \mathcal{H}) \qquad \lim_{k\to\infty} \langle \boldsymbol{u}, \boldsymbol{x}^{[k]} - \widehat{\boldsymbol{x}}\rangle = 0.$$

It is denoted by $\boldsymbol{x}^{[k]} \rightharpoonup \widehat{\boldsymbol{x}}$.

Remark: In a finite dimensional Hilbert space, strong and weak convergences are equivalent.

Banach-Picard theorem

$\mathfrak{T}\colon \mathcal{H} \to \mathcal{H}$ is $\omega-$**Lipschitz continuous** for some $\omega > 0$ if

$$(\forall \boldsymbol{x} \in \mathcal{H})(\forall \boldsymbol{u} \in \mathcal{H}) \qquad \|\mathfrak{T}\boldsymbol{x} - \mathfrak{T}\boldsymbol{u}\| \leq \omega \|\boldsymbol{x} - \boldsymbol{u}\|.$$

$\mathfrak{T}$ is **nonexpansive** if it is $1-$Lipschitz continuous.

Banach-Picard theorem:

Let $\omega \in [0, 1)$, $\mathfrak{T}\colon \mathcal{H} \to \mathcal{H}$ be a $\omega-$Lipschitz continuous operator, and $\boldsymbol{x}^{[0]} \in \mathcal{H}$.
Set

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \mathfrak{T}\boldsymbol{x}^{[k]}.$$

Then, $\operatorname{Fix} \mathfrak{T} = \{\widehat{\boldsymbol{x}}\}$ for some $\widehat{\boldsymbol{x}} \in \mathcal{H}$ and we have

$$(\forall k \in \mathbb{N}) \quad \|\boldsymbol{x}^{[k]} - \widehat{\boldsymbol{x}}\| \leq \omega^k \|\boldsymbol{x}_0 - \widehat{\boldsymbol{x}}\|.$$

Moreover, $(\boldsymbol{x}^{[k]})_{k \in \mathbb{N}}$ converges strongly to $\widehat{\boldsymbol{x}}$ with linear convergence rate ω.

Averaged nonexpansive operator

An operator $\mathfrak{T}\colon \mathcal{H} \to \mathcal{H}$ is **$\mu-$averaged nonexpansive** for some $\mu \in (0,1]$ if, for every $\boldsymbol{x} \in \mathcal{H}$ and $\boldsymbol{u} \in \mathcal{H}$,

$$\|\mathfrak{T}\boldsymbol{x} - \mathfrak{T}\boldsymbol{u}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{u}\|^2 - \left(\frac{1-\mu}{\mu}\right) \|(\mathrm{Id} - \mathfrak{T})\boldsymbol{x} - (\mathrm{Id} - \mathfrak{T})\boldsymbol{u}\|^2$$

$\mathfrak{T}$ is **firmly nonexpansive** if it is $1/2-$averaged.

$\mathfrak{T}$ is **nonexpansive** if and only if $\mathfrak{T}$ is $1-$averaged.

Theorem:
Let $\mu \in (0,1)$, let $\mathfrak{T}\colon \mathcal{H} \to \mathcal{H}$ be a $\mu-$averaged nonexpansive operator such that $\mathrm{Fix}\,\mathfrak{T} \neq \varnothing$, and let $\boldsymbol{x}^{[0]} \in \mathcal{H}$.
Set $(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \mathfrak{T}\boldsymbol{x}^{[k]}$.
Then $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converges weakly to a point in $\mathrm{Fix}\mathfrak{T}$.

Nonlinear operators

Properties of f	$\mathfrak{T}$	ω-Lipschitz	μ-averaged
f convex ∇f β-Lipschitz	$\mathrm{Id} - \tau\nabla f$ $\tau \in (0, 2\beta^{-1})$	$\omega = 1$	$\mu = \frac{\tau\beta}{2}$
f ρ-strongly convex ∇f β-Lipschitz	$\mathrm{Id} - \tau\nabla f$ $\tau \in (0, 2\beta^{-1})$	$\omega = \max\{(1-\tau\rho), (\tau\beta - 1)\}$	$\mu = \frac{1+\omega}{2}$
$f \in \Gamma_0(\mathcal{H})$	$\mathrm{prox}_{\tau f}$ $\tau > 0$	$\omega = 1$	$\mu = \frac{1}{2}$
f ρ-strongly convex	$\mathrm{prox}_{\tau f}$ $\tau > 0$	$\omega = (1+\tau\rho)^{-1}$	$\mu = \frac{1+\omega}{2}$

Proximal algorithms

- **Minimisation problem** :

$$\boxed{\widehat{\boldsymbol{x}} \in \underset{\boldsymbol{x}}{\operatorname{Argmin}} \;\; f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})}$$

with f_1 and f_2 either diff. with Lipschitz gradient or proximable.

- **Design of a sequence of the form:**

$$(\forall k \in \mathbb{N}) \qquad \boldsymbol{x}^{[k+1]} = \mathfrak{T}\boldsymbol{x}^{[k]},$$

Gradient descent	$\mathfrak{T} = \mathrm{Id} - \tau(\nabla f_1 + \nabla f_2)$
Proximal point	$\mathfrak{T} = \mathrm{prox}_{\tau(f_1+f_2)}$
Forward-Backward	$\mathfrak{T} = \mathrm{prox}_{\tau f_2}(\mathrm{Id} - \tau\nabla f_1)$
Peaceman-Rachford	$\mathfrak{T} = (2\mathrm{prox}_{\tau f_2} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau f_1} - \mathrm{Id})$
Douglas-Rachford	$\mathfrak{T} = \mathrm{prox}_{\tau f_2}(2\mathrm{prox}_{\tau f_1} - \mathrm{Id}) + \mathrm{Id} - \mathrm{prox}_{\tau f_1}$

Primal algorithms

FB algorithm $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}}\Big\{f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})\Big\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau\nabla f_1)$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_2}(\boldsymbol{x}^{[k]} - \tau\nabla f_1(\boldsymbol{x}^{[k]}))$.

FB algorithm $\widehat{x} \in \operatorname{Argmin}_{x\in\mathcal{H}}\Big\{ f(x) = f_1(x) + f_2(x) \Big\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau\nabla f_1)$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x^{[k+1]} = \operatorname{prox}_{\tau f_2}(x^{[k]} - \tau\nabla f_1(x^{[k]}))$.
- Roots in projected gradient method [Levitin, 1966] when $f_2 = \iota_C$ for some closed convex set C.
- If $f_2 = 0$, gradient descent algorithm
- if $f_1 = 0$, proximal point algorithm.

FB algorithm $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} \Big\{ f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \Big\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau\nabla f_1)$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_2}(\boldsymbol{x}^{[k]} - \tau\nabla f_1(\boldsymbol{x}^{[k]}))$.
- For every $\tau > 0$, $\operatorname{zer}(\nabla f_1 + \partial f_2) = \operatorname{Fix}\mathfrak{T}$.

  Proof:

$$\begin{aligned} \boldsymbol{x} \in \operatorname{Fix}\mathfrak{T} &\Leftrightarrow (\operatorname{Id} - \tau\nabla f_1)\boldsymbol{x} \in (\operatorname{Id} + \tau\partial f_2)\boldsymbol{x} \\ &\Leftrightarrow 0 \in \nabla f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x}). \end{aligned}$$

FB algorithm $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} \Big\{ f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \Big\}$

Objective: Let $f_1 \colon \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$, $\mathfrak{T} := \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau \nabla f_1)$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_2}(\boldsymbol{x}^{[k]} - \tau \nabla f_1(\boldsymbol{x}^{[k]}))$.
- For every $\tau > 0$, $\operatorname{zer}(\nabla f_1 + \partial f_2) = \operatorname{Fix} \mathfrak{T}$.
- $\operatorname{prox}_{\tau f_2}(\operatorname{Id} - \tau \nabla f_1)$ is μ-averaged nonexpansive where $\mu = \frac{\mu_1 + \mu_2 - 2\mu_1\mu_2}{1 - \mu_1\mu_2}$ with $\mu_2 = \tau\beta/2$ and $\mu_1 = 1/2$ leading to $\mu = \frac{1}{2 - \tau\beta/2} \in (0,1)$ and $\tau < 2/\beta$.

FB algorithm $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} \Big\{ f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \Big\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau \nabla f_1)$

Theorem [Combettes & Wajs, 2005]:

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be a sequence generated by the FB algorithm. Let $\tau \in (0, 2\beta^{-1})$. Then

- $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converges to a minimiser $\widehat{\boldsymbol{x}}$ of f (if there exists one)
- $\big(f(\boldsymbol{x}^{[k]})\big)_{k\in\mathbb{N}}$ is a non-increasing sequence converging to $f(\widehat{\boldsymbol{x}})$.

FB algorithm $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} \left\{ f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \right\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau \nabla f_1)$

Theorem [Briceno-Arias & Pustelnik, 2005]

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be a sequence generated by the FB algorithm.

- Suppose that f_1 is ρ-strongly convex, and $\tau \in (0, 2\beta^{-1})$. Then $\mathfrak{T}$ is $\omega(\tau)$−Lipschitz continuous with $\omega(\tau) := \max\left\{|1-\tau\rho|, |1-\tau\beta|\right\} \in (0,1)$

  In particular, the minimum is achieved at $\tau^* = \frac{2}{\rho+\beta}$ and $\omega(\tau^*) = \frac{\beta-\rho}{\beta+\rho}$

FB algorithm $\widehat{\boldsymbol{x}} \in \text{Argmin}_{\boldsymbol{x}\in\mathcal{H}} \Big\{ f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \Big\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \text{prox}_{\tau f_2} \circ (\text{Id} - \tau \nabla f_1)$

Theorem [Briceno-Arias & Pustelnik, 2005]

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be a sequence generated by the FB algorithm.

- Suppose that f_1 is ρ-strongly convex, and $\tau \in (0, 2\beta^{-1})$. Then $\mathfrak{T}$ is $\omega(\tau)-$Lipschitz continuous with $\omega(\tau) := \max\big\{|1-\tau\rho|, |1-\tau\beta|\big\} \in (0,1)$

  In particular, the minimum is achieved at $\tau^* = \frac{2}{\rho+\beta}$ and $\omega(\tau^*) = \frac{\beta-\rho}{\beta+\rho}$

- Suppose that f_2 is ρ-strongly convex, and $\tau \in (0, 2\beta^{-1})$. Then $\mathfrak{T}$ is $\omega(\tau)-$Lipschitz continuous with $\omega(\tau) := \frac{1}{1+\tau\rho} \in (0,1)$

  In particular, the minimum is achieved at $\tau^* = \frac{2}{\beta}$ and $\omega(\tau^*) = \frac{\beta}{\beta+2\rho}$

FB algorithm $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} \Big\{ f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \Big\}$

Objective: Let $f_1 : \mathcal{H} \to \mathbb{R}$ a convex, proper and β-Lipschitz differentiable function and $f_2 \in \Gamma_0(\mathcal{H})$.
We set, for some $\tau > 0$, $\mathfrak{T} := \mathrm{prox}_{\tau f_2} \circ (\mathrm{Id} - \tau \nabla f_1)$

- Convergence may be slow in practice...
  - ☛ Use Nesterov acceleration (*inertia*/*momentum*)
  - ☛ Use second order information (*preconditioning*)
  - ☛ Use multilevel strategy
- What if $\mathrm{prox}_{\gamma_k f_2}$ does not have a closed form?
  - ☛ Use sub-iterations (e.g. dual FB algorithm)
  - ☛ Use more advanced methods (e.g. primal-dual algorithms)

Acceleration via Inertia

What is inertia?

Goal: Inertia aims to use information from the **previous iterate(s)** $(\boldsymbol{x}^{[k']})_{k'\leq k}$ to build the next iterate $\boldsymbol{x}^{[k+1]}$.

Why? Use memory to go faster!

For FB we have

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \mathfrak{T}_k(\boldsymbol{x}^{[k]}) \text{ where } \mathfrak{T}_k = \operatorname{prox}_{\tau f_2} \circ (\operatorname{Id} - \tau \nabla f_1)$$

Introducing inertia would lead to

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \widetilde{\mathfrak{T}}_k(\boldsymbol{x}^{[1]}, \ldots, \boldsymbol{x}^{[k]})$$

Question: How to choose $\widetilde{\mathfrak{T}}_k$?

Remark: In general $\widetilde{\mathfrak{T}}_k$ only depends on $(\boldsymbol{x}^{[k]}, \boldsymbol{x}^{[k-1]})$ to avoid memory issues

Particular case: Inertia for GD algorithm

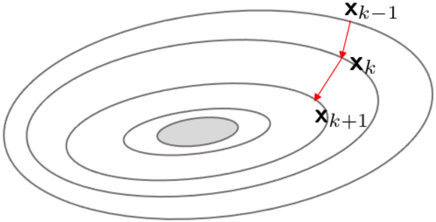
Let $f_2 \equiv 0$. In this case $\operatorname{prox}_{f_2} = \mathrm{Id}$.

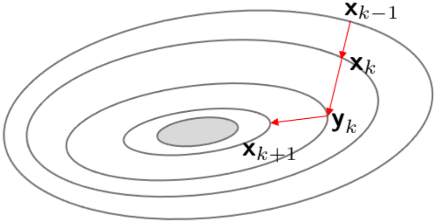
The *path* taken by the iterates $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ is determined by the opposite of the gradient direction:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \tau_k \nabla f_1(\boldsymbol{x}^{[k]})$$

Acceleration: *Nesterov-type accelerated GD algorithm* [Nesterov, 1983]

$$(\forall k \in \mathbb{N}) \qquad \begin{aligned} \boldsymbol{x}^{[k+1]} &= \boldsymbol{y}^{[k]} - \tau \nabla f_1(\boldsymbol{y}^{[k]}) \quad \text{with } \tau \in (0, 1/\beta] \\ \boldsymbol{y}_{k+1} &= \boldsymbol{x}^{[k+1]} + \alpha_k(\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}) \end{aligned}$$





Particular case: Inertia for GD algorithm

Let $f_2 \equiv 0$. In this case $\mathrm{prox}_{f_2} = \mathrm{Id}$.

The *path* taken by the iterates $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ is determined by the opposite of the gradient direction:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \tau_k \nabla f_1(\boldsymbol{x}^{[k]})$$

Acceleration: *Nesterov-type accelerated GD algorithm* [Nesterov, 1983]

$$(\forall k \in \mathbb{N}) \qquad \begin{aligned} \boldsymbol{x}^{[k+1]} &= \boldsymbol{y}^{[k]} - \tau \nabla f_1(\boldsymbol{y}^{[k]}) \quad \text{with } \tau \in (0, 1/\beta] \\ \boldsymbol{y}_{k+1} &= \boldsymbol{x}^{[k+1]} + \alpha_k(\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}) \end{aligned}$$

- Each iteration takes nearly the same computational cost as GD
- **not** a *descent* method (i.e. we may not have $f_1(\boldsymbol{x}^{[k+1]}) \leq f_1(\boldsymbol{x}^{[k]})$)

Inertial FB

Inertial FB

For $k = 0, 1, \dots$

Let $\gamma_k \in (0, 1/\beta]$

$$\boldsymbol{x}^{[k+1]} = \mathrm{prox}_{\tau_k f_2}\Big(\boldsymbol{y}^{[k]} - \tau_k \nabla f_1(\boldsymbol{y}^{[k]})\Big)$$

$$\boldsymbol{y}^{[k+1]} = \boldsymbol{x}^{[k+1]} + \alpha_k(\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})$$

- [Beck & Teboulle, 2009]
  Adopt the inertia (momentum) strategy proposed by Nesterov

$$\alpha_k = \frac{\theta_k - 1}{\theta_{k+1}} \quad \text{with} \quad \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}, \theta_1 = 0$$

Convergence rate for Inertial FB

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by FB iterations with $\tau \in (0, \beta^{-1}]$.
$(f(\boldsymbol{x}^{[k]}))_{k\in\mathbb{N}}$ converges to $f(\widehat{\boldsymbol{x}})$ at the rate $O(1/k)$:

$$f(\boldsymbol{x}^{[k]}) - f(\widehat{\boldsymbol{x}}) \leq \frac{\beta}{2k}\|\boldsymbol{x}^{[0]} - \widehat{\boldsymbol{x}}\|^2$$

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by Inertial FB.
$(f(\boldsymbol{x}^{[k]}))_{k\in\mathbb{N}}$ converges to $f(\widehat{\boldsymbol{x}})$ at the rate $O(1/k^2)$:

$$f(\boldsymbol{x}^{[k]}) - f(\widehat{\boldsymbol{x}}) \leq \frac{2\beta}{(k+1)^2}\|\boldsymbol{x}^{[0]} - \widehat{\boldsymbol{x}}\|^2$$

Convergence rate for Inertial FB

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by FB iterations with $\tau \in (0, \beta^{-1}]$.
$(f(\boldsymbol{x}^{[k]}))_{k\in\mathbb{N}}$ converges to $f(\widehat{\boldsymbol{x}})$ at the rate $O(1/k)$:

$$f(\boldsymbol{x}^{[k]}) - f(\widehat{\boldsymbol{x}}) \leq \frac{\beta}{2k}\|\boldsymbol{x}^{[0]} - \widehat{\boldsymbol{x}}\|^2$$

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by Inertial FB.
$(f(\boldsymbol{x}^{[k]}))_{k\in\mathbb{N}}$ converges to $f(\widehat{\boldsymbol{x}})$ at the rate $O(1/k^2)$:

$$f(\boldsymbol{x}^{[k]}) - f(\widehat{\boldsymbol{x}}) \leq \frac{2\beta}{(k+1)^2}\|\boldsymbol{x}^{[0]} - \widehat{\boldsymbol{x}}\|^2$$

- Improved iteration complexity
- (Almost) same computational complexity per iteration as FB
- Issue: Does the sequence $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converge?

Convergence rate for Inertial FB

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by FB iterations with $\tau \in (0, \beta^{-1}]$.
$(f(\boldsymbol{x}^{[k]}))_{k\in\mathbb{N}}$ converges to $f(\widehat{\boldsymbol{x}})$ at the rate $O(1/k)$:

$$f(\boldsymbol{x}^{[k]}) - f(\widehat{\boldsymbol{x}}) \leq \frac{\beta}{2k} \|\boldsymbol{x}^{[0]} - \widehat{\boldsymbol{x}}\|^2$$

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by Inertial FB.
$(f(\boldsymbol{x}^{[k]}))_{k\in\mathbb{N}}$ converges to $f(\widehat{\boldsymbol{x}})$ at the rate $O(1/k^2)$:

$$f(\boldsymbol{x}^{[k]}) - f(\widehat{\boldsymbol{x}}) \leq \frac{2\beta}{(k+1)^2} \|\boldsymbol{x}^{[0]} - \widehat{\boldsymbol{x}}\|^2$$

Let $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ be generated by Inertial FB with Chambolle-Dossal rule $\alpha_k = \frac{\theta_k - 1}{\theta_{k+1}}$ with $\theta_{k+1} = \left(\frac{k+a}{a}\right)^d$ with $d \in (0,1]$ and $a > \max\{1, (2d)^{1/d}\}$
Then the sequence $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converges weakly to a minimiser of f.

Duality

[Komodakis & Pesquet, 2015]

Minimization problem

Find

$$\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$$

- $f_1 : \mathbb{R}^N \to \mathbb{R}$ is convex and β-Lipschitz differentiable
- $f_2 \in \Gamma_0(\mathcal{H})$
- $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H},\mathcal{G})$: space of linear continuous operators

Use FB algorithm ?

For $k = 0, 1, \ldots$

$$\left\lfloor \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau(f_2+g\circ\mathbf{W})}\left(\boldsymbol{x}^{[k]} - \tau\nabla f_1(\boldsymbol{x}^{[k]})\right)\right.$$

How to compute $\operatorname{prox}_{\tau(f_2+g\circ\mathbf{W})}$?

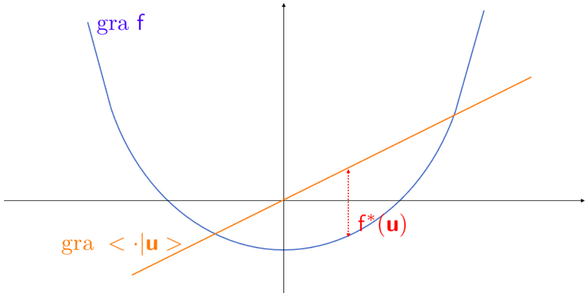
➤ Use primal-dual methods

Conjugate function

The conjugate of a function $f\colon \mathcal{H} \to]-\infty,+\infty]$ is the **convex** function f^* defined as

$$f^*\colon \begin{array}{ccc} \mathcal{H} & \to & [-\infty,+\infty] \\ \boldsymbol{u} & \mapsto & \sup_{\boldsymbol{x}\in\mathcal{H}} \langle \boldsymbol{x} \mid \boldsymbol{u}\rangle - f(\boldsymbol{x}) \end{array}$$

Graphical illustration: $f^*(\boldsymbol{u})$ is the supremum of the signed vertical distance between the graph of f and that of the linear functional $\langle \cdot \mid \boldsymbol{u}\rangle$



Conjugate function

The conjugate of a function $f \colon \mathcal{H} \to]-\infty, +\infty]$ is the **convex** function f^* defined as

$$
\begin{array}{rccl}
f^* \colon & \mathcal{H} & \to & [-\infty, +\infty] \\
& \boldsymbol{u} & \mapsto & \sup_{\boldsymbol{x} \in \mathcal{H}} \langle \boldsymbol{x} \mid \boldsymbol{u} \rangle - f(\boldsymbol{x})
\end{array}
$$

Example :

- $f = \frac{1}{2}\| \cdot \|^2 \Rightarrow f^* = \frac{1}{2}\| \cdot \|^2$.

Conjugate function

The conjugate of a function $f\colon \mathcal{H} \to]-\infty, +\infty]$ is the **convex** function f^* defined as

$$\begin{array}{rccl} f^*\colon & \mathcal{H} & \to & [-\infty, +\infty] \\ & \boldsymbol{u} & \mapsto & \sup_{\boldsymbol{x} \in \mathcal{H}} \langle \boldsymbol{x} \mid \boldsymbol{u} \rangle - f(\boldsymbol{x}) \end{array}$$

Moreau-Fenchel theorem

Let $\mathcal{H}$ be a Hilbert space and $f\colon \mathcal{H} \to (-\infty, +\infty]$ be a proper function.

$$f \text{ is l.s.c. and convex} \Leftrightarrow f^{**} = f.$$

In general, f^{**} is the lower semi-continuous convex enveloppe of f.

Conjugate: properties

Fenchel-Young inequality: If f is proper, then

1. $\big(\forall(\boldsymbol{x},\boldsymbol{u})\in\mathcal{H}^2\big) \qquad f(\boldsymbol{x})+f^*(\boldsymbol{u})\geq\langle\boldsymbol{x}\mid\boldsymbol{u}\rangle$

2. $\big(\forall(\boldsymbol{x},\boldsymbol{u})\in\mathcal{H}^2\big) \qquad \boldsymbol{u}\in\partial f(\boldsymbol{x})\Leftrightarrow \quad f(\boldsymbol{x})+f^*(\boldsymbol{u})=\langle\boldsymbol{x}\mid\boldsymbol{u}\rangle.$

If $f\in\Gamma_0(\mathcal{H})$, then

$$\big(\forall(\boldsymbol{x},\boldsymbol{u})\in\mathcal{H}^2\big) \qquad \boldsymbol{u}\in\partial f(\boldsymbol{x})\ \Leftrightarrow\ \boldsymbol{x}\in\partial f^*(\boldsymbol{u}).$$

Equivalently, $\partial f^*=(\partial f)^{-1}$.

Conjugate: Moreau decomposition

Moreau decomposition formula: Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$.

$$(\forall \boldsymbol{x} \in \mathcal{H}) \qquad \operatorname{prox}_{\gamma f^*}(\boldsymbol{x}) = \boldsymbol{x} - \gamma \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1} \boldsymbol{x}).$$

Conjugate: Moreau decomposition

Moreau decomposition formula: Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$.

$$(\forall \boldsymbol{x} \in \mathcal{H}) \qquad \operatorname{prox}_{\gamma f^*}(\boldsymbol{x}) = \boldsymbol{x} - \gamma \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1}\boldsymbol{x}).$$

Example: If C is a nonempty closed convex set of $\mathcal{H}$, its indicator function is

$$(\forall \boldsymbol{x} \in \mathcal{H}) \quad \iota_C(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \in C \\ +\infty & \text{otherwise.} \end{cases}$$

The conjugate of ι_C is the **support function** of C: $(\forall \boldsymbol{u} \in \mathcal{H}) \quad \iota_C^*(\boldsymbol{u}) = \sup_{\boldsymbol{x} \in \mathcal{H}} \langle \boldsymbol{u} \mid \boldsymbol{x} \rangle$

and $\operatorname{prox}_{\iota_C^*} = \mathrm{Id} - \operatorname{proj}_C$.

Special case: $\mathcal{H} = \mathbb{R}^N$, $C = [-\delta, \delta]^N$ with $\delta > 0$, $\iota_C^* = \delta \| \cdot \|_1$

$\Rightarrow \operatorname{prox}_{\iota_C^*} = \mathrm{Id} - \operatorname{proj}_{[-\delta,\delta]^N}$: soft-thresholding with threshold δ

Fenchel-Rockafellar duality

Primal problem

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let $f\colon \mathcal{H} \to (-\infty, +\infty]$, $g\colon \mathcal{G} \to (-\infty, +\infty]$. Let $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.
We want to

$$\underset{x \in \mathcal{H}}{\text{minimize}}\ f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}).$$

Dual problem

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let $f\colon \mathcal{H} \to (-\infty, +\infty]$, $g\colon \mathcal{G} \to (-\infty, +\infty]$. Let $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.
We want to

$$\underset{\boldsymbol{u} \in \mathcal{G}}{\text{minimize}}\ f^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u}).$$

Fenchel-Rockafellar duality

Weak duality

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let f be a proper fonction from $\mathcal{H}$ to $(-\infty, +\infty]$, g be a proper function from $\mathcal{G}$ to $(-\infty, +\infty]$, and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Let

$$\mu = \inf_{\boldsymbol{x} \in \mathcal{H}} f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) \quad \text{and} \quad \mu^* = \inf_{\boldsymbol{u} \in \mathcal{G}} f^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u}).$$

We have $\mu \geq -\mu^*$. If $\mu \in \mathbb{R}$, $\mu + \mu^*$ is called the **duality gap**.

Fenchel-Rockafellar duality

Weak duality

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let f be a proper fonction from $\mathcal{H}$ to $(-\infty, +\infty]$, g be a proper function from $\mathcal{G}$ to $(-\infty, +\infty]$, and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Let

$$\mu = \inf_{\boldsymbol{x} \in \mathcal{H}} f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) \quad \text{and} \quad \mu^* = \inf_{\boldsymbol{u} \in \mathcal{G}} f^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u}).$$

We have $\mu \geq -\mu^*$. If $\mu \in \mathbb{R}$, $\mu + \mu^*$ is called the **duality gap**.

Proof: According to Fenchel-Young inequality,

$$f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) + f^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u}) \geq \langle \boldsymbol{x} \mid -\mathbf{W}^*\boldsymbol{u} \rangle + \langle \mathbf{W}\boldsymbol{x} \mid \boldsymbol{u} \rangle = 0.$$

Fenchel-Rockafellar duality

Strong duality

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$, and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.
If $\mathrm{int}(\mathrm{dom} g) \cap \mathbf{W}(\mathrm{dom} f) \neq \emptyset$ or $\mathrm{dom} g \cap \mathrm{int}\big(\mathbf{W}(\mathrm{dom} f)\big) \neq \emptyset$, then

$$\mu = \inf_{\boldsymbol{x} \in \mathcal{H}} f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) = -\min_{\boldsymbol{u} \in \mathcal{G}} f^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u}) = -\mu^* .$$

Fenchel-Rockafellar duality

Duality theorem (2)

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$, and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

- If there exists $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}}) \in \mathcal{H} \times \mathcal{G}$ such that $-\mathbf{W}^*\widehat{\boldsymbol{u}} \in \partial f(\widehat{\boldsymbol{x}})$ and $\mathbf{W}\widehat{\boldsymbol{x}} \in \partial g^*(\widehat{\boldsymbol{u}})$, then $\widehat{\boldsymbol{x}}$ (resp. $\widehat{\boldsymbol{u}}$) is a solution to the primal (resp. dual) problem.

If $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}}) \in \mathcal{H} \times \mathcal{G}$ is such that $-\mathbf{W}^*\widehat{\boldsymbol{u}} \in \partial f(\widehat{\boldsymbol{x}})$ and $\mathbf{W}\widehat{\boldsymbol{x}} \in \partial g^*(\widehat{\boldsymbol{u}})$, then $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}})$ is called a **Kuhn-Tucker point**.

Forward-Backward iterations in the dual

Dual FB algorithm

Let $\boldsymbol{z} \in \mathcal{H}$, $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^N}{\operatorname{Argmin}} \; f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u} \in \mathbb{R}^M}{\operatorname{Argmin}} \; \widetilde{f^*}(\boldsymbol{z} - \mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

Dual FB algorithm

Let $\boldsymbol{z} \in \mathcal{H}$, $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^N}{\operatorname{Argmin}} \; f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u} \in \mathbb{R}^M}{\operatorname{Argmin}} \; \widetilde{f^*}(\boldsymbol{z} - \mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

- $\widetilde{f^*}$ is the Moreau enveloppe of f^*
- $\widetilde{f^*}$ is differentiable and $\nabla \widetilde{f^*} = \mathrm{Id} - \operatorname{prox}_{f^*} = \operatorname{prox}_f$ 1-Lipschitz continuous
- Use FB on the dual problem!

Dual FB algorithm

Let $\boldsymbol{z} \in \mathcal{H}$, $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^N}{\operatorname{Argmin}} \; f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u} \in \mathbb{R}^M}{\operatorname{Argmin}} \; \widetilde{f^*}(\boldsymbol{z} - \mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

Choose $\boldsymbol{u}_0 \in \mathbb{R}^M$ and $\tau \in (0, 2/\|\mathbf{W}\|^2)$.

For $k = 0, 1, \ldots$

$$
\left\lfloor
\begin{array}{l}
\boldsymbol{x}^{[k]} = \operatorname{prox}_f\left(\boldsymbol{z} - \mathbf{W}^*\boldsymbol{u}^{[k]}\right) \\
\boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\tau g^*}\left(\boldsymbol{u}^{[k]} + \tau \mathbf{W}\boldsymbol{x}^{[k]}\right)
\end{array}
\right.
$$

[Combettes, Dung & Vũ, 2011]

The sequence $(\boldsymbol{u}^{[k]})_{k\in\mathbb{N}}$ converges weakly to a solution to the dual problem $\widehat{\boldsymbol{u}}$.

The sequence $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converges strongly to a solution to the primal problem $\widehat{\boldsymbol{x}} = \operatorname{prox}_f(\boldsymbol{z} - \mathbf{W}^*\widehat{\boldsymbol{u}})$.

ADMM

Augmented Lagrangian method

ADMM algorithm (*Alternating-direction method of multipliers*)
⇒ **Lagrangian interpretation**

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{minimize}}\; f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) \quad \Leftrightarrow \quad \underset{\substack{\boldsymbol{x}\in\mathcal{H},\boldsymbol{u}\in\mathcal{G}\\ \mathbf{W}\boldsymbol{x}=\boldsymbol{u}}}{\operatorname{minimize}}\; f(\boldsymbol{x}) + g(\boldsymbol{u})$$

- Lagrange function : $\mathcal{L}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = f(\boldsymbol{x}) + g(\boldsymbol{u}) + \langle \boldsymbol{v} \mid \mathbf{W}\boldsymbol{x} - \boldsymbol{u}\rangle$

⇒ $\boldsymbol{v} \in \mathcal{G}$ is the Lagrange multiplier.

Augmented Lagrangian method

ADMM algorithm (*Alternating-direction method of multipliers*)
$\Rightarrow$ **Lagrangian interpretation**

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\text{minimize}}\ f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) \quad \Leftrightarrow \quad \underset{\substack{\boldsymbol{x}\in\mathcal{H},\boldsymbol{u}\in\mathcal{G}\\ \mathbf{W}\boldsymbol{x}=\boldsymbol{u}}}{\text{minimize}}\ f(\boldsymbol{x}) + g(\boldsymbol{u})$$

- Lagrange function: $\mathcal{L}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = f(\boldsymbol{x}) + g(\boldsymbol{u}) + \langle \boldsymbol{v} \mid \mathbf{W}\boldsymbol{x} - \boldsymbol{u}\rangle$

$\Rightarrow$ $\boldsymbol{v} \in \mathcal{G}$ is the Lagrange multiplier.

- Idea: iterations for finding a saddle point $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}})$:

$$(\forall n \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \operatorname{Argmin}\ \mathcal{L}(\cdot, \boldsymbol{u}^{[k]}, \boldsymbol{v}^{[k]}) \\ \boldsymbol{u}^{[k+1]} \in \operatorname{Argmin}\ \mathcal{L}(\boldsymbol{x}^{[k]}, \cdot, \boldsymbol{v}^{[k]}) \\ \boldsymbol{v}^{[k+1]} \text{ such that } \mathcal{L}(\boldsymbol{x}^{[k]}, \boldsymbol{u}^{[k+1]}, \boldsymbol{v}^{[k+1]}) \geq \mathcal{L}(\boldsymbol{x}^{[k]}, \boldsymbol{u}^{[k+1]}, \boldsymbol{v}^{[k]}). \end{cases}$$

But **convergence not guaranteed in general !**

Augmented Lagrangian method

ADMM algorithm (*Alternating-direction method of multipliers*)
⇒ **Lagrangian interpretation**

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\text{minimize}}\; f(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x}) \quad \Leftrightarrow \quad \underset{\substack{\boldsymbol{x}\in\mathcal{H},\boldsymbol{u}\in\mathcal{G}\\ \mathbf{W}\boldsymbol{x}=\boldsymbol{u}}}{\text{minimize}}\; f(\boldsymbol{x}) + g(\boldsymbol{u})$$

- Lagrange function: $\mathcal{L}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = f(\boldsymbol{x}) + g(\boldsymbol{u}) + \langle \boldsymbol{v} \mid \mathbf{W}\boldsymbol{x} - \boldsymbol{u}\rangle$

⇒ $\boldsymbol{v} \in \mathcal{G}$ is the Lagrange multiplier.

- Solution: introduce an **Augmented Lagrange function**:

$$\widetilde{\mathcal{L}}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{w}) = f(\boldsymbol{x}) + g(\boldsymbol{u}) + \gamma\,\langle \boldsymbol{w} \mid \mathbf{W}\boldsymbol{x} - \boldsymbol{u}\rangle + \frac{\gamma}{2}\|\mathbf{W}\boldsymbol{x} - \boldsymbol{u}\|^2$$

⇒ The Lagrange multiplier is $\boldsymbol{v} = \gamma\boldsymbol{w}$ with $\gamma > 0$.

Alternating-direction method of multipliers

Algorithm for finding a saddle point:

$$(\forall k \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{Argmin}} \ \widetilde{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{y}^{[k]}, \boldsymbol{w}^{[k]}) \\ \boldsymbol{y}^{[k+1]} \in \underset{\boldsymbol{y}\in\mathcal{G}}{\operatorname{Argmin}} \ \widetilde{\mathcal{L}}(\boldsymbol{x}^{[k]}, \boldsymbol{y}, \boldsymbol{w}^{[k]}) \\ \boldsymbol{w}^{[k+1]} \text{ such that } \widetilde{\mathcal{L}}(\boldsymbol{x}^{[k]}, \boldsymbol{y}^{[k+1]}, \boldsymbol{w}^{[k+1]}) \geq \widetilde{\mathcal{L}}(\boldsymbol{x}^{[k]}, \boldsymbol{y}^{[k+1]}, \boldsymbol{w}^{[k]}). \end{cases}$$

By performing a gradient ascent on the Lagrange multiplier,

$$(\forall k \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{Argmin}} \ f(\boldsymbol{x}) + \gamma \left\langle \boldsymbol{w}^{[k]} \mid \mathbf{W}\boldsymbol{x} - \boldsymbol{y}^{[k]} \right\rangle + \frac{\gamma}{2}\|\mathbf{W}\boldsymbol{x} - \boldsymbol{y}^{[k]}\|^2 \\ \boldsymbol{y}^{[k+1]} \in \underset{\boldsymbol{y}\in\mathcal{G}}{\operatorname{Argmin}} \ g(\boldsymbol{y}) + \gamma \left\langle \boldsymbol{w}^{[k]} \mid \mathbf{W}\boldsymbol{x}^{[k]} - \boldsymbol{y} \right\rangle + \frac{\gamma}{2}\|\mathbf{W}\boldsymbol{x}^{[k]} - \boldsymbol{y}\|^2 \\ \boldsymbol{w}^{[k+1]} = \boldsymbol{w}^{[k]} + \frac{1}{\gamma}\nabla_{\boldsymbol{w}}\widetilde{\mathcal{L}}(\boldsymbol{x}^{[k]}, \boldsymbol{y}^{[k+1]}, \boldsymbol{w}^{[k]}) \end{cases}$$

$$\Leftrightarrow \quad (\forall k \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{Argmin}} \ \frac{1}{2}\left\|\mathbf{W}\boldsymbol{x} - \boldsymbol{y}^{[k]} + \boldsymbol{w}^{[k]}\right\|^2 + \frac{1}{\gamma} f(\boldsymbol{x}) \\ \boldsymbol{y}^{[k+1]} = \operatorname{prox}_{\frac{g}{\gamma}}\left(\boldsymbol{w}^{[k]} + \mathbf{W}\boldsymbol{x}^{[k]}\right) \\ \boldsymbol{w}^{[k+1]} = \boldsymbol{w}^{[k]} + \mathbf{W}\boldsymbol{x}^{[k]} - \boldsymbol{y}^{[k+1]}. \end{cases}$$

Augmented Lagrange method

ADMM algorithm (*Alternating-direction method of multipliers*)

Let $f \in \Gamma_0(\mathcal{H})$ et $g \in \Gamma_0(\mathcal{G})$. Let $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ such that $\mathbf{W}^*\mathbf{W}$ is an isomorphism and $\gamma > 0$. Let

$$(\forall k \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \underset{\boldsymbol{x} \in \mathcal{H}}{\operatorname{Argmin}} \ \frac{1}{2} \left\| \mathbf{W}\boldsymbol{x} - \boldsymbol{y}^{[k]} + \boldsymbol{w}^{[k]} \right\|^2 + \frac{1}{\gamma} f(\boldsymbol{x}) \\ \boldsymbol{s}^{[k]} = \mathbf{W}\boldsymbol{x}^{[k]} \\ \boldsymbol{y}^{[k+1]} = \operatorname{prox}_{\frac{g}{\gamma}} \left(\boldsymbol{w}^{[k]} + \boldsymbol{s}^{[k]} \right) \\ \boldsymbol{w}^{[k+1]} = \boldsymbol{w}^{[k]} + \boldsymbol{s}^{[k]} - \boldsymbol{y}^{[k+1]}. \end{cases}$$

Augmented Lagrange method

ADMM algorithm (*Alternating-direction method of multipliers*)

Let $f \in \Gamma_0(\mathcal{H})$ et $g \in \Gamma_0(\mathcal{G})$. Let $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ such that $\mathbf{W}^*\mathbf{W}$ is an isomorphism and $\gamma > 0$.
Let

$$(\forall k \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \underset{\boldsymbol{x} \in \mathcal{H}}{\operatorname{Argmin}} \ \frac{1}{2} \left\| \mathbf{W}\boldsymbol{x} - \boldsymbol{y}^{[k]} + \boldsymbol{w}^{[k]} \right\|^2 + \frac{1}{\gamma} f(\boldsymbol{x}) \\ \boldsymbol{s}^{[k]} = \mathbf{W}\boldsymbol{x}^{[k]} \\ \boldsymbol{y}^{[k+1]} = \operatorname{prox}_{\frac{g}{\gamma}} \left(\boldsymbol{w}^{[k]} + \boldsymbol{s}^{[k]} \right) \\ \boldsymbol{w}^{[k+1]} = \boldsymbol{w}^{[k]} + \boldsymbol{s}^{[k]} - \boldsymbol{y}^{[k+1]}. \end{cases}$$

We assume that $\operatorname{int}(\operatorname{dom} g) \cap \mathbf{W}(\operatorname{dom} f) \neq \varnothing$ or $\operatorname{dom} g \cap \operatorname{int}\big(\mathbf{W}(\operatorname{dom} f)\big) \neq \varnothing$ and that $\operatorname{Argmin}(f + g \circ \mathbf{W}) \neq \varnothing$.

- $\boldsymbol{x}^{[k]} \rightharpoonup \widehat{\boldsymbol{x}} \in \operatorname{Argmin}(f + g \circ \mathbf{W})$
- $\gamma \boldsymbol{w}^{[k]} \rightharpoonup \widehat{\boldsymbol{v}} \in \operatorname{Argmin}(f^* \circ (-\mathbf{W}^*) + g^*)$.

Augmented Lagrange method

ADMM algorithm (*Alternating-direction method of multipliers*)

Let $f \in \Gamma_0(\mathcal{H})$ et $g \in \Gamma_0(\mathcal{G})$. Let $\mathbf{W} \in \mathcal{B}(\mathcal{H},\mathcal{G})$ such that $\mathbf{W}^*\mathbf{W}$ is an isomorphism and $\gamma > 0$. Let

$$(\forall k \in \mathbb{N}) \qquad \begin{cases} \boldsymbol{x}^{[k]} \in \underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{Argmin}} \ \frac{1}{2}\left\|\mathbf{W}\boldsymbol{x} - \boldsymbol{y}^{[k]} + \boldsymbol{w}^{[k]}\right\|^2 + \frac{1}{\gamma} f(\boldsymbol{x}) \\ \boldsymbol{s}^{[k]} = \mathbf{W}\boldsymbol{x}^{[k]} \\ \boldsymbol{y}^{[k+1]} = \operatorname{prox}_{\frac{g}{\gamma}}\left(\boldsymbol{w}^{[k]} + \boldsymbol{s}^{[k]}\right) \\ \boldsymbol{w}^{[k+1]} = \boldsymbol{w}^{[k]} + \boldsymbol{s}^{[k]} - \boldsymbol{y}^{[k+1]}. \end{cases}$$

We assume that $\operatorname{int}(\operatorname{dom} g) \cap \mathbf{W}(\operatorname{dom} f) \neq \varnothing$ or $\operatorname{dom} g \cap \operatorname{int}\big(\mathbf{W}(\operatorname{dom} f)\big) \neq \varnothing$ and that $\operatorname{Argmin}\,(f + g \circ \mathbf{W}) \neq \varnothing$.

- $\boldsymbol{x}^{[k]} \rightharpoonup \widehat{\boldsymbol{x}} \in \operatorname{Argmin}\,(f + g \circ \mathbf{W})$

≡ **Douglas-Rachford for the dual problem**

Primal-dual forward-backward iterations

Primal-dual problem formulation

Let $f_1 \in \Gamma_0(\mathcal{H})$, $f_2 \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} \in \underset{\boldsymbol{x} \in \mathcal{H}}{\operatorname{Argmin}} \; f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u} \in \mathcal{G}}{\operatorname{Argmin}} \; (f_1 + f_2)^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

Primal-dual problem formulation

Let $f_1 \in \Gamma_0(\mathcal{H})$, $f_2 \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} \in \underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{Argmin}} \; f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u}\in\mathcal{G}}{\operatorname{Argmin}} \; (f_1 + f_2)^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

Lagrangian-like formulation: Another formulation of the Primal-Dual problem is to combine them into the search of a **saddle point of the function:**

$$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}}) \in \underset{\boldsymbol{x}\in\mathcal{H}}{\operatorname{Argmin}} \, \max_{\boldsymbol{u}\in\mathcal{G}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) - g^*(\boldsymbol{u}) + \langle \mathbf{W}\boldsymbol{x}, \boldsymbol{u} \rangle$$

Primal-dual problem formulation

Let $f_1 \in \Gamma_0(\mathcal{H})$, $f_2 \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} \in \underset{\boldsymbol{x} \in \mathcal{H}}{\operatorname{Argmin}} \; f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u} \in \mathcal{G}}{\operatorname{Argmin}} \; (f_1 + f_2)^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

Lagrangian-like formulation: Another formulation of the Primal-Dual problem is to combine them into the search of a **saddle point of the function:**

$$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}}) \in \underset{\boldsymbol{x} \in \mathcal{H}}{\operatorname{Argmin}} \max_{\boldsymbol{u} \in \mathcal{G}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) - g^*(\boldsymbol{u}) + \langle \mathbf{W}\boldsymbol{x}, \boldsymbol{u} \rangle$$

Karush-Kuhn-Tucker conditions: Assume that $\operatorname{dom} g \cap \mathbf{W}(\operatorname{dom} f) \neq \emptyset$ and f_2 differentiable.

$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{u}}) \in \mathcal{H} \times \mathcal{G}$ is a solution to the Primal-Dual problem if and only if

$$\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \in \begin{pmatrix} \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{pmatrix}$$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W} \widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

Multiply by $\tau > 0$ the first equation and $\sigma > 0$ the second equation:

$$\begin{cases} -\tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) \in \tau \partial f_1(\widehat{\boldsymbol{x}}) \\ \sigma \mathbf{W}\widehat{\boldsymbol{x}} \in \sigma \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

Multiply by $\tau > 0$ the first equation and $\sigma > 0$ the second equation:

$$\begin{cases} -\tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) \in \tau \partial f_1(\widehat{\boldsymbol{x}}) \\ \sigma \mathbf{W}\widehat{\boldsymbol{x}} \in \sigma \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

These equations are equivalent to

$$\begin{cases} \widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) - \widehat{\boldsymbol{x}} \in \tau \partial f_1(\widehat{\boldsymbol{x}}) \\ \\ \widehat{\boldsymbol{u}} + \sigma \mathbf{W}\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{u}} \in \sigma \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

Multiply by $\tau > 0$ the first equation and $\sigma > 0$ the second equation:

$$\begin{cases} -\tau\big(\mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) \in \tau\partial f_1(\widehat{\boldsymbol{x}}) \\ \sigma\mathbf{W}\widehat{\boldsymbol{x}} \in \sigma\partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

These equations are equivalent to

$$\begin{cases} \widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) - \widehat{\boldsymbol{x}} \in \tau\partial f_1(\ \widehat{\boldsymbol{x}}\ ) \\ \\ \widehat{\boldsymbol{u}} + \sigma\mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}) - \widehat{\boldsymbol{u}} \in \sigma\partial g^*(\ \widehat{\boldsymbol{u}}\ ) \end{cases}$$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

Multiply by $\tau > 0$ the first equation and $\sigma > 0$ the second equation:

$$\begin{cases} -\tau\big(\mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) \in \tau\partial f_1(\widehat{\boldsymbol{x}}) \\ \sigma\mathbf{W}\widehat{\boldsymbol{x}} \in \sigma\partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

These equations are equivalent to

$$\begin{cases} \underbrace{\widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big)}_{\overline{\boldsymbol{x}}} - \underbrace{\widehat{\boldsymbol{x}}}_{\overline{\mathbf{p}}} \in \tau\partial f(\underbrace{\widehat{\boldsymbol{x}}}_{\overline{\mathbf{p}}}) & \rightsquigarrow \operatorname{prox}_{\tau f_1} \\ \underbrace{\widehat{\boldsymbol{u}} + \sigma\mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}})}_{\overline{\boldsymbol{x}}} - \underbrace{\widehat{\boldsymbol{u}}}_{\overline{\mathbf{p}}} \in \sigma\partial g^*(\underbrace{\widehat{\boldsymbol{u}}}_{\overline{\mathbf{p}}}) & \rightsquigarrow \operatorname{prox}_{\sigma g^*} \end{cases}$$

Prox characterisation: $\overline{\boldsymbol{x}} - \overline{\mathbf{p}} \in \gamma\partial\psi(\overline{\mathbf{p}}) \Leftrightarrow \overline{\mathbf{p}} = \operatorname{prox}_{\gamma\psi}(\overline{\boldsymbol{x}})$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

Multiply by $\tau > 0$ the first equation and $\sigma > 0$ the second equation:

$$\begin{cases} -\tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) \in \tau\partial f_1(\widehat{\boldsymbol{x}}) \\ \sigma\mathbf{W}\widehat{\boldsymbol{x}} \in \sigma\partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

These equations are equivalent to

$$\begin{cases} \underbrace{\widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big)}_{\overline{\boldsymbol{x}}} - \underbrace{\widehat{\boldsymbol{x}}}_{\overline{\mathbf{p}}} \in \tau\partial f(\underbrace{\widehat{\boldsymbol{x}}}_{\overline{\mathbf{p}}}) \quad \rightsquigarrow \operatorname{prox}_{\tau f_1} \\ \underbrace{\widehat{\boldsymbol{u}} + \sigma\mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}})}_{\overline{\boldsymbol{x}}} - \underbrace{\widehat{\boldsymbol{u}}}_{\overline{\mathbf{p}}} \in \sigma\partial g^*(\underbrace{\widehat{\boldsymbol{u}}}_{\overline{\mathbf{p}}}) \quad \rightsquigarrow \operatorname{prox}_{\sigma g^*} \end{cases}$$

$$\Leftrightarrow \begin{cases} \widehat{\boldsymbol{x}} = \operatorname{prox}_{\tau f_1}\Big(\widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big)\Big) \\ \widehat{\boldsymbol{u}} = \operatorname{prox}_{\sigma g^*}\Big(\widehat{\boldsymbol{u}} + \sigma\mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}})\Big) \end{cases}$$

From KKT to fixed-point equations...

KKT:

$$\begin{cases} \mathbf{0} \in \partial f_1(\widehat{\boldsymbol{x}}) + \mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}}) \\ \mathbf{0} \in -\mathbf{W}\widehat{\boldsymbol{x}} + \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

Multiply by $\tau > 0$ the first equation and $\sigma > 0$ the second equation:

$$\begin{cases} -\tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big) \in \tau \partial f_1(\widehat{\boldsymbol{x}}) \\ \sigma \mathbf{W}\widehat{\boldsymbol{x}} \in \sigma \partial g^*(\widehat{\boldsymbol{u}}) \end{cases}$$

These equations are equivalent to

$$\begin{cases} \underbrace{\widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big)}_{\overline{\boldsymbol{x}}} - \underbrace{\widehat{\boldsymbol{x}}}_{\overline{\mathbf{p}}} \in \tau \partial f(\underbrace{\widehat{\boldsymbol{x}}}_{\overline{\mathbf{p}}}) & \rightsquigarrow \operatorname{prox}_{\tau f_1} \\ \underbrace{\widehat{\boldsymbol{u}} + \sigma \mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}})}_{\overline{\boldsymbol{x}}} - \underbrace{\widehat{\boldsymbol{u}}}_{\overline{\mathbf{p}}} \in \sigma \partial g^*(\underbrace{\widehat{\boldsymbol{u}}}_{\overline{\mathbf{p}}}) & \rightsquigarrow \operatorname{prox}_{\sigma g^*} \end{cases}$$

$$\Leftrightarrow \begin{cases} \widehat{\boldsymbol{x}} = \operatorname{prox}_{\tau f_1}\Big(\widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^* \widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big)\Big) \\ \widehat{\boldsymbol{u}} = \operatorname{prox}_{\sigma g^*}\Big(\widehat{\boldsymbol{u}} + \sigma \mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}})\Big) \end{cases} \qquad \rightsquigarrow \textit{Fixed-point equations}$$

Fixed-point algorithm

From the fixed-point equations:

$$\begin{cases} \widehat{\boldsymbol{x}} = \operatorname{prox}_{\tau f_1}\Big(\widehat{\boldsymbol{x}} - \tau\big(\mathbf{W}^*\widehat{\boldsymbol{u}} + \nabla f_2(\widehat{\boldsymbol{x}})\big)\Big) \\ \widehat{\boldsymbol{u}} = \operatorname{prox}_{\sigma g^*}\Big(\widehat{\boldsymbol{u}} + \sigma\mathbf{W}(2\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}})\Big) \end{cases}$$

we derive a fixed-point algorithm:

For $k = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}\Big(\boldsymbol{x}^{[k]} - \tau\big(\mathbf{W}^*\boldsymbol{u}^{[k]} + \nabla f_2(\boldsymbol{x}^{[k]})\big)\Big) \\ \boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}\Big(\boldsymbol{u}^{[k]} + \sigma\mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\Big) \end{array} \right.$$

Remark:

- Algorithm known as the Condat-Vũ algorithm

Step-size and convergence of Condat-Vũ algorithm

Let $f_1 \in \Gamma_0(\mathcal{H})$, $f_2 \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $\mathbf{W} \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

Primal problem: $\widehat{\boldsymbol{x}} \in \underset{\boldsymbol{x} \in \mathcal{H}}{\operatorname{Argmin}} \; f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Dual problem: $\widehat{\boldsymbol{u}} \in \underset{\boldsymbol{u} \in \mathcal{G}}{\operatorname{Argmin}} \; (f_1 + f_2)^*(-\mathbf{W}^*\boldsymbol{u}) + g^*(\boldsymbol{u})$

Choose $\tau > 0$ and $\sigma > 0$ such that $\frac{1}{\tau} - \sigma\|\mathbf{W}\|^2 > \frac{\beta}{2}$ with ∇f_2 β-Lipschitz gradient.

For $k = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}\Big(\boldsymbol{x}^{[k]} - \tau\big(\nabla f_2(\boldsymbol{x}^{[k]}) + \mathbf{W}^*\boldsymbol{u}^{[k]}\big)\Big) \\ \boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}\Big(\boldsymbol{u}^{[k]} + \sigma\mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\Big) \end{array} \right.$$

[Vũ, 2013][Condat, 2013]

The sequence $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converges weakly to a solution to the primal problem.

The sequence $(\boldsymbol{u}^{[k]})_{k\in\mathbb{N}}$ converges weakly to a solution to the dual problem.

Particular cases

CONDAT-VŨ ALGORITHM: [Vũ, 2013][Condat, 2013]

PROBLEM: Find $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Choose $\tau > 0$ and $\sigma > 0$ such that $\frac{1}{\tau} - \sigma\|\mathbf{W}\|^2 > \frac{\beta}{2}$ with f_2 β-Lipschitz.

For $k = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}\Big(\boldsymbol{x}^{[k]} - \tau\big(\nabla f_2(\boldsymbol{x}^{[k]}) + \mathbf{W}^*\boldsymbol{u}^{[k]}\big)\Big) \\ \boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}\Big(\boldsymbol{u}^{[k]} + \sigma\mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\Big) \end{array} \right.$$

Particular cases

Condat-Vũ algorithm: [Vũ, 2013][Condat, 2013]

Problem: Find $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Choose $\tau > 0$ and $\sigma > 0$ such that $\frac{1}{\tau} - \sigma\|\mathbf{W}\|^2 > \frac{\beta}{2}$ with f_2 β-Lipschitz.

For $k = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}\left(\boldsymbol{x}^{[k]} - \tau\left(\nabla f_2(\boldsymbol{x}^{[k]}) + \mathbf{W}^*\boldsymbol{u}^{[k]}\right)\right) \\ \boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}\left(\boldsymbol{u}^{[k]} + \sigma\mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\right) \end{array}\right.$$

Chambolle-Pock (CP) algorithm: $f_2 \equiv 0$ [Chambolle & Pock, 2011]

Problem: Find $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} f_1(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Choose $\tau > 0$ and $\sigma > 0$ such that $\sigma\tau\|\mathbf{W}\|^2 < 1$.

For $k = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}(\boldsymbol{x}^{[k]} - \tau\mathbf{W}^*\boldsymbol{u}^{[k]}) \\ \boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}(\boldsymbol{u}^{[k]} + \sigma\mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})) \end{array}\right.$$

Particular cases

Condat-Vũ algorithm: [Vũ, 2013][Condat, 2013]

Problem: Find $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

Choose $\tau > 0$ and $\sigma > 0$ such that $\frac{1}{\tau} - \sigma\|\mathbf{W}\|^2 > \frac{\beta}{2}$ with f_2 β-Lipschitz.

For $k = 0, 1, \ldots$

$$\boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}\Big(\boldsymbol{x}^{[k]} - \tau\big(\nabla f_2(\boldsymbol{x}^{[k]}) + \mathbf{W}^*\boldsymbol{u}^{[k]}\big)\Big)$$

$$\boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}\Big(\boldsymbol{u}^{[k]} + \sigma\mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\Big)$$

Douglas-Rachford (DR) algorithm: $f_2 \equiv 0$, $\mathbf{W} = \mathrm{Id}$ and $\tau = 1/\sigma$

Problem: Find $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathcal{H}} f_1(\boldsymbol{x}) + g(\boldsymbol{x})$

Choose $\sigma > 0$.

For $k = 0, 1, \ldots$

$$\boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\sigma^{-1} f_1}(\boldsymbol{s}_k)$$

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \boldsymbol{x}^{[k+1]} - \operatorname{prox}_{\sigma^{-1} g}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{s}_k)$$

CP algorithm and strong convexity $\widehat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathbb{R}^N} f_1(\boldsymbol{x}) + g(\mathbf{W}\boldsymbol{x})$

CHAMBOLLE-POCK ALGORITHM: [Chambolle & Pock, 2011]

Choose $\tau > 0$ and $\sigma > 0$ such that $\sigma\tau\|\mathbf{W}\|^2 < 1$.
For $k = 0, 1, \ldots$

$$\boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau f_1}\left(\boldsymbol{x}^{[k]} - \tau \mathbf{W}^* \boldsymbol{u}^{[k]}\right)$$
$$\boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma g^*}\left(\boldsymbol{u}^{[k]} + \sigma \mathbf{W}(2\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\right)$$

ACCELERATED VERSION: f_1 ρ-strongly convex [Chambolle & Pock, 2011]

Choose $\tau_0 > 0$ and $\sigma_0 > 0$ such that $\sigma_0\tau_0\|\mathbf{W}\|^2 < 1$.
For $k = 0, 1, \ldots$

$$\boldsymbol{x}^{[k+1]} = \operatorname{prox}_{\tau_k f_1}\left(\boldsymbol{x}^{[k]} - \tau_k \mathbf{W}^* \boldsymbol{u}^{[k]}\right)$$
$$\alpha_k = (1 + 2\rho\tau_k)^{-1/2}$$
$$\tau_{k+1} = \alpha_k \tau_k$$
$$\sigma_k = \sigma_k \alpha_k^{-1/2}$$
$$\boldsymbol{y}^{[k+1]} = \boldsymbol{x}^{[k+1]} + \alpha_k(\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})$$
$$\boldsymbol{u}^{[k+1]} = \operatorname{prox}_{\sigma_{k+1} g^*}\left(\boldsymbol{u}^{[k]} + \sigma \mathbf{W}\boldsymbol{y}^{[k+1]}\right)$$

Optimization algorithms

Forward-Backward	$f_1 + f_2$	f_1 grad. Lipschitz prox_{f_2}	[Combettes,Wajs,2005]
ISTA	$f_1 + f_2$	f_1 grad. Lipschitz $f_2 = \lambda\|\cdot\|_1$	[Daubechies et al, 2003]
Douglas-Rachford	$f_1 + f_2$	prox_{f_1} prox_{f_2}	[Combettes,Pesquet, 2007]
PPXA	$\sum_i f_i$	prox_{f_i}	[Combettes,Pesquet, 2008]
PPXA+	$\sum_i g_i \circ \mathbf{W}_i$	prox_{g_i} $(\sum_{i=1}^m \mathbf{W}_i^* \mathbf{W}_i)^{-1}$	[Pesquet, Pustelnik, 2012]
ADMM	$f + g \circ \mathbf{W}$	prox_f $(\mathbf{W}^*\mathbf{W})^{-1}$	[Eckstein, Yao, 2015]
Chambolle-Pock	$f + g \circ \mathbf{W}$	prox_f prox_g	[Chambolle, Pock, 2011]
Condat-Vũ	$f_1 + f_2 + g \circ \mathbf{W}$	prox_f prox_g f_2 grad. Lipschitz	[Condat, 2013][Vũ, 2013]