# On Decomposing the Proximal Map

Yao-Liang Yu

University of Waterloo

Splitting Algorithms, Modern Operator Theory, and Applications
Oaxaca, Mexico
September 19, 2017

## Table of Contents

1 Motivation

2 Setup

3 A Naive Sufficient Condition

4 Three Case Studies

5 More Examples

# Regularized loss minimization

Generic form for many ML problems:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \ell(\boldsymbol{w}) + f(\boldsymbol{w})$$

- $\ell$ is the loss function;
- $f$ is the regularizer, usually a (semi)norm;

Special interest:

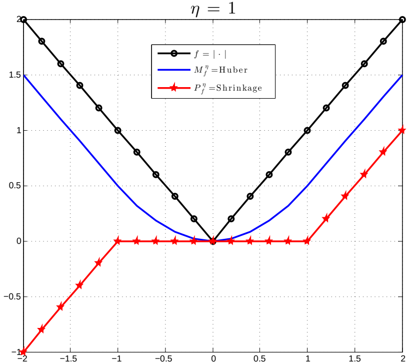
- sparsity;
- computational efficiency.

# Moreau envelop and proximal map

## Definition (Moreau'65)

$$\mathsf{M}_f(\boldsymbol{y}) = \min_{\boldsymbol{w}} \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{y}\|^2 + f(\boldsymbol{w})$$

$$\mathsf{P}_f(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}} \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{y}\|^2 + f(\boldsymbol{w})$$



# Proximal gradient (Fukushima & Mine'81)

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \ell(\boldsymbol{w}) + f(\boldsymbol{w})$$

1. $\boldsymbol{y}_t = \boldsymbol{w}_t - \eta \nabla \ell(\boldsymbol{w}_t)$;
2. $\boldsymbol{w}_{t+1} = \mathsf{P}_{\eta f}(\boldsymbol{y}_t)$.

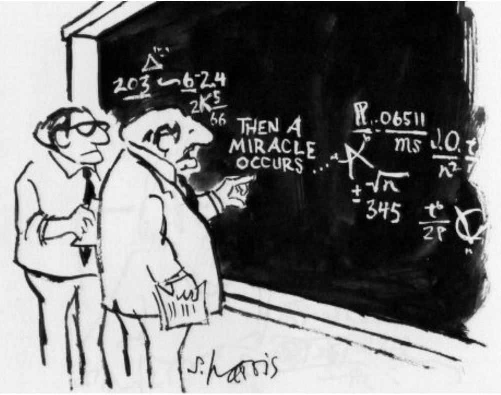
For $f = \|\cdot\|_1$, obtain the shrinkage operator

$$[\mathsf{P}_{\|\cdot\|_1}(\boldsymbol{y})]_i = \operatorname{sign}(y_i)(|y_i| - 1)_+.$$

- guaranteed convergence, can be accelerated;
- generalization of projected gradient: $f = \iota_C$;
- reveals the sparsity-inducing property.

Refs: Combettes & Wajs'05; Beck & Teboulle'09; Duchi & Singer'09; Nesterov'13; etc.

# Then A Miracle Occurs...




"I think you should be more explicit here in step two."

from *What's so Funny about Science?* by Sidney Harris (1977)

Step 2: $\mathsf{P}_f(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}\|^2 + f(\boldsymbol{w})$

# How to deal with sum?

- Typical structured sparse regularizers:

$$f(\boldsymbol{w}) = \sum_i f_i(\boldsymbol{w});$$

**Theorem (Parallel Sum)**

$$\mathsf{P}_{f+g} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id}).$$

- Not directly useful due to the inversion;
- Can numerically reduce to $\mathsf{P}_f$ and $\mathsf{P}_g$ (Combettes et al.'11);
- But a two-loop routine can be as slow as subgradient descent (Schmidt et.al'11; Villa et al.'13).

# Table of Contents

1 Motivation

2 Setup

3 A Naive Sufficient Condition

4 Three Case Studies

5 More Examples

# Two previous results

**Theorem (Friedman et al.'07)**

$$\mathsf{P}_{\|\cdot\|_1+\|\cdot\|_{\mathsf{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathsf{TV}}}, \quad \textit{where} \quad \|\boldsymbol{w}\|_{\mathsf{TV}} = \sum_{i=1}^{d-1} |w_i - w_{i+1}|.$$

**Theorem (Jenatton et al.'11)**

*Assuming the groups* $\{\mathrm{g}_i\}$ *form a laminar system* ($\mathrm{g}_i \cap \mathrm{g}_j \in \{\mathrm{g}_i, \mathrm{g}_j, \emptyset\}$)*, then, if appropriately ordered,*

$$\mathsf{P}_{\sum_{i=1}^k \|\cdot\|_{\mathrm{g}_i}} = \mathsf{P}_{\|\cdot\|_{\mathrm{g}_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{\mathrm{g}_k}},$$

*where* $\|\cdot\|_{\mathrm{g}_i}$ *is the restriction of* $\mathfrak{l}_p, p \in \{1, 2, \infty\}$ *to the group* $\mathrm{g}_i$*.*

# (Wild) Generalization

$$\mathsf{P}_{f+g} \stackrel{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \stackrel{?}{=} \mathsf{P}_g \circ \mathsf{P}_f.$$

## Product of Prox's

- Long line of work: von Neumann, Halperin, Amemiya and Ando, Stiles, Dye, Reich, Bruck, Tseng, Brézis and Lions, etc., etc.
- interest was in the asymptotic behaviour
- in some sense, we want one-step convergence of such algs

# Bad news

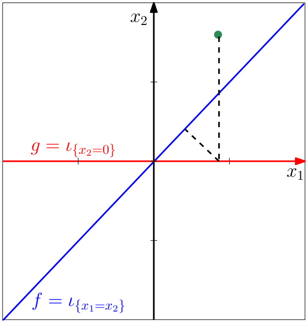
**Theorem**

*On the real line,* $\exists h$ *such that* $\mathsf{P}_h = \mathsf{P}_f \circ \mathsf{P}_g$.

- *Not necessarily* $h = f + g$, *though*

**Example (A simple counterexample)**

Consider $\mathbb{R}^2$, and let $f = \iota_{\{x_1 = x_2\}}, g = \iota_{\{x_2 = 0\}}$.




$$\mathsf{P}_f = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad \mathsf{P}_g = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

$$\text{But} \quad \mathsf{P}_f \circ \mathsf{P}_g = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \end{bmatrix}$$

no $h$ such that $\mathsf{P}_h = \mathsf{P}_f \circ \mathsf{P}_g$

# Table of Contents


1. Motivation
2. Setup
3. A Naive Sufficient Condition
4. Three Case Studies
5. More Examples

# Nevertheless

- Not possible to always have the decomposition — too ambitious
- More modest goal: decomposition to hold for certain functions
- Manipulating the optimality conditions:

$$\mathsf{P}_{f+g}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{w}\|^2 + (f+g)(\mathbf{w})$$
$$\mathsf{P}_{g}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{w}\|^2 + g(\mathbf{w})$$
$$\mathsf{P}_{f}(\mathsf{P}_{g}(\mathbf{z})) = \operatorname{argmin}_{\mathbf{w}} \tfrac{1}{2}\left\|\mathsf{P}_{g}(\mathbf{z})-\mathbf{w}\right\|^2 + f(\mathbf{w}).$$

**Theorem**

*A sufficient condition for* $\mathsf{P}_{f+g}(\mathbf{z}) = \mathsf{P}_{f}(\mathsf{P}_{g}(\mathbf{z}))$ *is*

$$\forall\, \mathbf{y} \in \operatorname{dom} g,\ \partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y}).$$

- "Proof" works as long as $f+g$ is convex

# Nevertheless

- Not possible to always have the decomposition — too ambitious
- More modest goal: decomposition to hold for certain functions
- Manipulating the optimality conditions:

$$\begin{aligned} \mathsf{P}_{f+g}(\boldsymbol{z}) - \boldsymbol{z} + \partial(f+g)(\mathsf{P}_{f+g}(\boldsymbol{z})) &\ni 0 \\ \mathsf{P}_g(\boldsymbol{z}) - \boldsymbol{z} + \partial g(\mathsf{P}_g(\boldsymbol{z})) &\ni 0 \\ \mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z})) - \mathsf{P}_g(\boldsymbol{z}) + \partial f(\mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z}))) &\ni 0. \end{aligned}$$

**Theorem**

*A sufficient condition for $\mathsf{P}_{f+g}(\boldsymbol{z}) = \mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z}))$ is*

$$\forall\, \boldsymbol{y} \in \operatorname{dom} g,\ \partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y}).$$

- "Proof" works as long as $f+g$ is convex

## Nevertheless

- Not possible to always have the decomposition — too ambitious
- More modest goal: decomposition to hold for certain functions
- Manipulating the optimality conditions:

$$\mathsf{P}_{f+g}(\boldsymbol{z}) - \boldsymbol{z} + \partial(f+g)(\mathsf{P}_{f+g}(\boldsymbol{z})) \ni 0$$
$$\mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z})) - \boldsymbol{z} + \partial g(\mathsf{P}_g(\boldsymbol{z})) + \partial f(\mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z}))) \ni 0.$$

**Theorem**

*A sufficient condition for* $\mathsf{P}_{f+g}(\boldsymbol{z}) = \mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z}))$ *is*

$$\forall\, \boldsymbol{y} \in \operatorname{dom} g,\ \partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y}).$$

- "Proof" works as long as $f+g$ is convex

# Nevertheless

- Not possible to always have the decomposition — too ambitious
- More modest goal: decomposition to hold for certain functions
- Manipulating the optimality conditions:

$$\mathsf{P}_{f+g}(\boldsymbol{z}) - \boldsymbol{z} + \partial(f+g)(\mathsf{P}_{f+g}(\boldsymbol{z})) \ni 0$$
$$\mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z})) - \boldsymbol{z} + \partial g(\mathsf{P}_g(\boldsymbol{z})) + \partial f(\mathsf{P}_f(\mathsf{P}_g(\boldsymbol{z}))) \ni 0.$$

**Theorem**

*A sufficient condition for* $\mathsf{P}_{f+g}(\boldsymbol{z}) = \mathsf{P}_f\big(\mathsf{P}_g(\boldsymbol{z})\big)$ *is*

$$\forall\ \boldsymbol{y} \in \operatorname{dom} g,\ \ \partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y}).$$

- "Proof" works as long as $f + g$ is convex

# The rest is easy

- Find $f$ and $g$ that clinch our sufficient condition.

# Recent Results

- More sufficient conditions in (Bauschke and Combettes, 2017)
- (Adly et al., 2017) removes any condition by re-defining one prox

# Table of Contents


1 Motivation

2 Setup

3 A Naive Sufficient Condition

4 Three Case Studies

5 More Examples

# "Trivialities"

**Theorem**

*Fix $f$. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ for all $g$ if and only if*

- $\dim(\mathcal{H}) \geq 2$; $f \equiv c$ *or* $f = \iota_{\{\boldsymbol{w}\}} + c$ *for some* $c \in \mathbb{R}$ *and* $\boldsymbol{w} \in \mathcal{H}$;
- $\dim(\mathcal{H}) = 1$ *and* $f = \iota_C + c$ *for some* $c \in \mathbb{R}$ *and set* $C$ *that is closed and convex.*

Asymmetry.

**Theorem**

*Fix $g$. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ for all $f$ if and only if $g$ is continuous affine.*

- Reassuring the impossibility to always have $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$;
- Still hope to get interesting results!

# Scaling Invariant $\Leftrightarrow$ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$$

$g$ positive homogeneous $\Leftrightarrow \forall \lambda > 0, \partial g(\lambda \boldsymbol{w}) = \partial g(\boldsymbol{w}) \Rightarrow \forall \boldsymbol{z}, \mathsf{P}_f(\boldsymbol{z}) \propto \boldsymbol{z}$

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x}+\boldsymbol{y}) \geq f(\boldsymbol{y})$;

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0, 1]$;

iv). $\boldsymbol{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

*If* $\dim(\mathcal{H}) = 1$, *only ii)* $\Longrightarrow$ *i) ceases to hold.*

# Scaling Invariant $\Leftrightarrow$ Positive Homogeneous

$$\partial g(\mathrm{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y})$$

$g$ positive homogeneous $\Leftrightarrow \forall \lambda > 0, \partial g(\lambda \mathbf{w}) = \partial g(\mathbf{w}) \Rightarrow \forall \mathbf{z}, \mathrm{P}_f(\mathbf{z}) \propto \mathbf{z}$

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $x \perp y$, $f(x + y) \geq f(y)$;

iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathrm{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv). $0 \in \operatorname{dom} f$ *and* $\mathrm{P}_{f+\kappa} = \mathrm{P}_f \circ \mathrm{P}_\kappa$ *for all positive homogeneous* $\kappa$.

*If* $\dim(\mathcal{H}) = 1$, *only* $ii) \implies i)$ *ceases to hold.*

# Scaling Invariant $\Leftrightarrow$ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$$

$g$ positive homogeneous $\Leftrightarrow \forall \lambda > 0, \partial g(\lambda \boldsymbol{w}) = \partial g(\boldsymbol{w}) \Rightarrow \forall \boldsymbol{z}, \mathsf{P}_f(\boldsymbol{z}) \propto \boldsymbol{z}$

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). *$f = h(\|\cdot\|)$ for some increasing function $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;*

ii). *For all perpendicular $x \perp y$, $f(x + y) \geq f(y)$;*

iii). *For all $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ for some $\lambda_{\boldsymbol{z}} \in [0, 1]$;*

iv). *$0 \in \operatorname{dom} f$ and $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ for all positive homogeneous $g$.*

# Scaling Invariant ⇔ Positive Homogeneous

$$\partial g(\mathsf{P}_f(y)) \supseteq \partial g(y)$$

$g$ positive homogeneous $\Leftrightarrow \forall \lambda > 0, \partial g(\lambda \boldsymbol{w}) = \partial g(\boldsymbol{w}) \Rightarrow \forall \boldsymbol{z}, \mathsf{P}_f(\boldsymbol{z}) \propto \boldsymbol{z}$

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$*;*

ii). *For all perpendicular* $x \perp y$, $f(x + y) \geq f(y)$*;*

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0, 1]$*;*

iv). $0 \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$*.*

*If* $\dim(\mathcal{H}) = 1$*, only ii)* $\Longrightarrow$ *i) ceases to hold.*

# Scaling Invariant ⇔ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$$

$g$ positive homogeneous ⇔ $\forall \lambda > 0, \partial g(\lambda \boldsymbol{w}) = \partial g(\boldsymbol{w}) \Rightarrow \forall \boldsymbol{z}, \mathsf{P}_f(\boldsymbol{z}) \propto \boldsymbol{z}$

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). *$f = h(\|\cdot\|)$ for some increasing function $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;*

ii). *For all perpendicular $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x} + \boldsymbol{y}) \geq f(\boldsymbol{y})$;*

iii). *For all $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ for some $\lambda_{\boldsymbol{z}} \in [0, 1]$;*

iv). *$\mathbf{0} \in \mathrm{dom}\, f$ and $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ for all positive homogeneous $\kappa$.*

*If $\dim(\mathcal{H}) = 1$, only ii) $\implies$ i) ceases to hold.*

# Scaling Invariant $\Leftrightarrow$ Positive Homogeneous

$$\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$$

$g$ positive homogeneous $\Leftrightarrow \forall \lambda > 0, \partial g(\lambda \boldsymbol{w}) = \partial g(\boldsymbol{w}) \Rightarrow \forall \boldsymbol{z}, \mathsf{P}_f(\boldsymbol{z}) \propto \boldsymbol{z}$

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x} + \boldsymbol{y}) \geq f(\boldsymbol{y})$;

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \mathrm{dom}\, f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

*If* $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

# Some Implications

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x}+\boldsymbol{y}) \geq f(\boldsymbol{y})$;

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0,1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

*If* $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

**i) $\iff$ ii)**

- Characterizing representer theorem (Dinuzzo & Schölkopf'12)

$$\operatorname{argmin} \ell(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle, \ldots, \langle \boldsymbol{w}, \boldsymbol{x}_n \rangle) + f(\boldsymbol{w}) \in \operatorname{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

# Some Implications

## Theorem

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x} + \boldsymbol{y}) \geq f(\boldsymbol{y})$;

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \mathrm{dom}\, f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

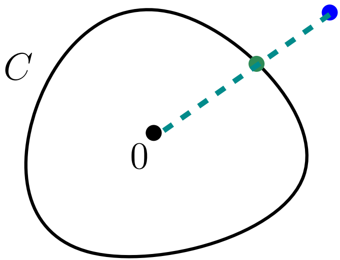
*If* $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

## i) $\iff$ ii)

- Characterizing representer theorem (Dinuzzo & Schölkopf'12)

  $$\operatorname{argmin} \ell(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle, \ldots, \langle \boldsymbol{w}, \boldsymbol{x}_n \rangle) + f(\boldsymbol{w}) \in \mathrm{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

# Characterizing the Ball



# Some Implications

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\mathbf{x} \perp \mathbf{y}$, $f(\mathbf{x} + \mathbf{y}) \geq f(\mathbf{y})$;

iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

*If* $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

**i) $\implies$ iv)**

$$\mathsf{P}_{\lambda\|\cdot\|^2+\kappa} = \mathsf{P}_{\lambda\|\cdot\|^2} \circ \mathsf{P}_\kappa = \tfrac{1}{\lambda+1}\mathsf{P}_\kappa$$

- Double shrinkage;
- $\kappa = \|\cdot\|_1$: Elastic net (Zou & Hastie'05);
- Adding an $\ell_2$-ish regularizer, computationally, is free.

# Some Implications

**Theorem**

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x} + \boldsymbol{y}) \geq f(\boldsymbol{y})$;

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \mathrm{dom}\, f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

*If* $\dim(\mathcal{H}) = 1$, *only ii)* $\implies$ *i) ceases to hold.*

**i) $\implies$ iv)**

$$\mathsf{P}_{\lambda\|\cdot\|^2+\kappa} = \mathsf{P}_{\lambda\|\cdot\|^2} \circ \mathsf{P}_\kappa = \frac{1}{\lambda+1}\mathsf{P}_\kappa$$

- Double shrinkage;
- $\kappa = \|\cdot\|_1$: Elastic net (Zou & Hastie'05);
- Adding an $\mathfrak{l}_2$-ish regularizer, computationally, is free.

# Some Implications

## Theorem

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\mathbf{x} \perp \mathbf{y}$, $f(\mathbf{x} + \mathbf{y}) \geq f(\mathbf{y})$;

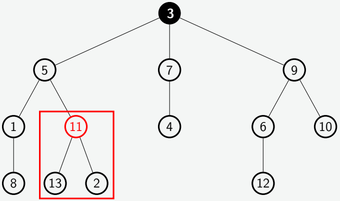
iii). *For all* $\mathbf{z} \in \mathcal{H}$, $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ *for some* $\lambda_{\mathbf{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

## i) $\implies$ iv)

Tree-structured group norms (Jenatton et al.'11)

$$\mathsf{P}_{\sum_i \|\cdot\|_{g_i}} = \mathsf{P}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{g_k}}.$$



# Some Implications

## Theorem

*Fix $f$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):*

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). *For all perpendicular* $\boldsymbol{x} \perp \boldsymbol{y}$, $f(\boldsymbol{x}+\boldsymbol{y}) \geq f(\boldsymbol{y})$;

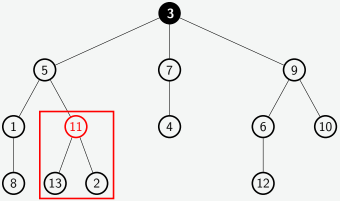
iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0,1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all positive homogeneous* $\kappa$.

## i) $\implies$ iv)

Tree-structured group norms (Jenatton et al.'11)

$$\mathsf{P}_{\sum_i \|\cdot\|_{g_i}} = \mathsf{P}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{g_k}}.$$



# Choquet Integral (*a.k.a.* Lovász Extension)

For an increasing set function $\mu : 2^{[d]} \to \mathbb{R}$:

$$g(\boldsymbol{w}) := \int_0^\infty \mu([\![\boldsymbol{w} \geq t]\!])\,\mathrm{d}t + \int_{-\infty}^0 [\mu([\![\boldsymbol{w} \geq t]\!]) - \mu([d])]\,\mathrm{d}t,$$

where we treat $w : \{1, \ldots, d\} \to \mathbb{R}$.

- $g$ is positive homogeneous
- $g(\boldsymbol{w} + \boldsymbol{z}) \neq g(\boldsymbol{w}) + g(\boldsymbol{z})$ in general
- $g(\boldsymbol{w} + \boldsymbol{z}) \leq g(\boldsymbol{w}) + g(\boldsymbol{z})$ iff $\mu$ is submodular:

$$\mu(A \cap B) + \mu(A \cup B) \leq \mu(A) + \mu(B)$$

- if $\forall i, j$, $(w_i - w_j)(z_i - z_j) \geq 0$, then $g(\boldsymbol{w} + \boldsymbol{z}) = g(\boldsymbol{w}) + g(\boldsymbol{z})$
- $\min_{A \subseteq [d]} \mu(A) = \min_{\boldsymbol{w} \in [0,1]} g(\boldsymbol{w})$.

# Further Properties of Choquet Integral

**Theorem**

*Let $g$ be the Choquet integral of some submodular function. If for all $i$ and $j$,*

- $(w_i - w_j)(z_i - z_j) \geq 0$, *then* $\partial g(\mathbf{w}) \cap \partial g(\mathbf{z}) \neq \emptyset$
- $w_i \geq w_j \implies z_i \geq z_j$, *then* $\partial g(\mathbf{w}) \subseteq \partial g(\mathbf{z})$.

**Theorem (Schmeidler'86)**

*If $g$ is comonotone additive and increasing/continuous, then $g$ is a Choquet integral of some set function.*

**TV is a Choquet integral**

$\|\mathbf{w}\|_{\mathrm{TV}} = \sum_{i=1}^{d-1} |w_i - w_{i+1}|$.

# Permutation Invariant $\Leftrightarrow$ Choquet Integral

$$\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$$

- For permutation-invariant $f$, recall

$$\mathsf{P}_f(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + f(\boldsymbol{x}).$$

By rearrangement inequality

$$y_i \geq y_j \implies [\mathsf{P}_f(\boldsymbol{y})]_i \geq [\mathsf{P}_f(\boldsymbol{y})]_j$$

**Theorem**

*Let $f$ be permutation invariant and $g$ be the Choquet integral of some submodular set function $\mu$. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.*

# Permutation Invariant $\Leftrightarrow$ Choquet Integral

$$\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$$

- For permutation-invariant $f$, recall

$$\mathsf{P}_f(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + f(\boldsymbol{x}).$$

By rearrangement inequality

$$y_i \geq y_j \implies [\mathsf{P}_f(\boldsymbol{y})]_i \geq [\mathsf{P}_f(\boldsymbol{y})]_j$$

**Theorem**

*Let $f$ be permutation invariant and $g$ be the Choquet integral of some submodular set function $\mu$. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.*

# Permutation Invariant $\Leftrightarrow$ Choquet Integral

$$\partial g(\mathsf{P}_f(y)) \supseteq \partial g(y)$$

- For permutation-invariant $f$, recall

$$\mathsf{P}_f(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + f(\boldsymbol{x}).$$

By rearrangement inequality

$$y_i \geq y_j \implies [\mathsf{P}_f(\boldsymbol{y})]_i \geq [\mathsf{P}_f(\boldsymbol{y})]_j$$

**Theorem**

*Let $f$ be permutation invariant and $g$ be the Choquet integral of some submodular set function $\mu$. Then, $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.*

# Some Implications

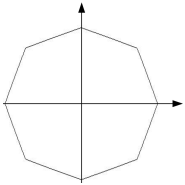
**Theorem**

*Let $f$ be permutation invariant and $g$ be the Choquet integral of some submodular set function. Then,* $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$.

- Special case $f = \|\cdot\|_1$ in (Bach'11);
- $\mathsf{P}_{\|\cdot\|_1 + \|\cdot\|_{\mathsf{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathsf{TV}}}$ (Friedman et al.'07);
- $\mathsf{P}_{\sum_{i=1}^k \|\cdot\|_{g_i}} = \mathsf{P}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{g_k}}$ (Jenatton et al.'11)

# Some Implications

$$\|\boldsymbol{w}\|_{\mathtt{oscar}} = \sum_{i<j} \max\{|w_i|, |w_j|\}.$$



- Feature grouping (Bondell & Reich'08)
- $\mathsf{P}_{\|\cdot\|_{\mathtt{oscar}}}$ in (Zhong & Kwok'11)

Let

$$\kappa_i(\boldsymbol{w}) := \sum_{j:j<i} \max\{|w_i|, |w_j|\}.$$

- $\|\boldsymbol{w}\|_{\mathtt{oscar}} = \sum_{i=2}^{d} \kappa_i(\boldsymbol{w})$
- $\mathsf{P}_{\|\cdot\|_{\mathtt{oscar}}} = \mathsf{P}_{\kappa_d} \circ \cdots \circ \mathsf{P}_{\kappa_2}$
- Given $\mathsf{P}_{\kappa_i}$, constant time for $\mathsf{P}_{\kappa_{i+1}}$.

# Table of Contents

1 Motivation

2 Setup

3 A Naive Sufficient Condition

4 Three Case Studies

5 More Examples

# Projection to intersection

**Theorem (Barty, Roy and Strugarek, MOR'07, Proposition 3.1)**

*Let* $L \cap C \neq \emptyset$*, where* $C$ *is a closed convex set and* $L$ *is a subspace. If* $\mathsf{P}_C(L) \subseteq L$*, then* $\mathsf{P}_{L \cap C} = \mathsf{P}_C \circ \mathsf{P}_L$*.*

$f = \iota_C$ and $g = \iota_L$, follows from $\partial g \equiv L^\perp$.

# One Solution for All, I

**Theorem (Chambolle and Darbon, 2009)**

*Let* $\varphi_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}, i = 1, \ldots, d,$ *be closed convex univariate functions and* $f : \mathbb{R}^d \to \mathbb{R}$ *be the Choquet integral of the set function* $\mu$. *Let*

$$\boldsymbol{u} \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \sum_{i=1}^{d} \varphi_i(w_i), \tag{1}$$

*whose existence is assumed. For any* $t \in \cap_i \operatorname{dom} \partial\varphi_i$, *consider the discrete problem:*

$$\min_{A \subseteq [d]} F(A) + \sum_{i \in A} \varphi_i'(t). \tag{2}$$

- *If for all* $i$, $\varphi_i'(t)$ *is the smallest element in the subdifferential* $\partial\varphi_i(t)$ *(existence assumed), then the set* $[\![\boldsymbol{u} \geq t]\!]$ *solves* (2).
- *If for all* $i$, $\varphi_i'(t)$ *is the largest element in the subdifferential* $\partial\varphi_i(t)$ *(existence assumed), then the set* $[\![\boldsymbol{u} > t]\!]$ *solves* (2).

# One Solution for All, II

## Theorem (extending (Barlow and Brunk, 1972))

*Let $f$ be univariate convex and differentiable, with the induced Bregman divergence $D_f(x, y) := f(x) - f(y) - f'(y)(x - y)$. For any Choquet integral $g$, the following problem*

$$\min_{\boldsymbol{x} \in \mathbb{R}^p} \quad \sum_{i=1}^{p} w_i D_f(x_i, y_i) + g(\boldsymbol{x}) \tag{3}$$

*can be solved in two steps:*

$$① \ \boldsymbol{z} = \operatorname*{argmin}_{\boldsymbol{x}} \frac{1}{2} \sum_{i=1}^{p} w_i (x_i - f'(y_i))^2 + g(\boldsymbol{x})$$

$$② \ \boldsymbol{y}^\star = (f')^{-1}(\boldsymbol{z}),$$

# Summary

- Posed the question: $\mathsf{P}_{f+g} \overset{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \overset{?}{=} \mathsf{P}_g \circ \mathsf{P}_f$;

- Presented a sufficient condition: $\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$;

- Identified two major cases;

- Immediately useful if plugged into splitting algs;