The Proximity Operator

Yao-Liang Yu
Machine Learning Department
Carnegie Melon University
Pittsburgh, PA, 15213, USA
yaoliang@cs.cmu.edu

February 26, 2014

Abstract

We present some basic properties of the proximity operator.

1 Notation

Our underlying universe is the (real) Hilbert space $\mathcal{H}$, equipped with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. For example, $\mathcal{H} = \mathbb{R}^d, \langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^d x_i y_i, \|\mathbf{x}\| = \sqrt{\sum_{i=1}^d x_i^2}$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$.

We will denote $C \subseteq \mathcal{H}$ as a nonempty, closed and convex set. Denote Γ_0 as the set of all functions $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ that are closed, proper (i.e., nonempty effective domain) and convex. All of our functions will be elements of Γ_0. We define the indicator function

$$\iota_C(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in C \\ \infty, & \text{otherwise} \end{cases}. \tag{1}$$

2 Historical Motivation

Fix any nonempty closed convex set C. We recall the projection operator:

$$\mathsf{P}_{\iota_C}(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{x} \in C} \tfrac{1}{2} \|\mathbf{z} - \mathbf{x}\|^2 . \tag{2}$$

We learned the following result in a linear algebra course:

Theorem 1 *Let $M \subseteq \mathcal{H}$ be a (closed) subspace and $M^\perp := \{\mathbf{y} \in \mathcal{H} : \forall \mathbf{x} \in M, \langle \mathbf{x}, \mathbf{y} \rangle = 0\}$ its orthogonal complement. Then the following are equivalent:*

- $\mathbf{z} = \mathbf{x} + \mathbf{y}, \mathbf{x} \in M, \mathbf{y} \in M^\perp$ *(implying $\langle \mathbf{x}, \mathbf{y} \rangle = 0$);*
- $\mathbf{x} = \mathsf{P}_{\iota_M}(\mathbf{z}), \mathbf{y} = \mathsf{P}_{\iota_{M^\perp}}(\mathbf{z})$.

A very natural and useful generalization of the above orthogonal decomposition is as follows:

Theorem 2 (Moreau'62) *Let $K \subseteq \mathcal{H}$ be a (closed) convex cone and $K^\circ := \{\mathbf{y} \in \mathcal{H} : \forall \mathbf{x} \in K, \langle \mathbf{x}, \mathbf{y} \rangle \leq 0\}$ its polar cone. Then the following are equivalent:*

- $\mathbf{z} = \mathbf{x} + \mathbf{y}, \mathbf{x} \in K, \mathbf{y} \in K^\circ$ *and $\langle \mathbf{x}, \mathbf{y} \rangle = 0$;*
- $\mathbf{x} = \mathsf{P}_{\iota_K}(\mathbf{z}), \mathbf{y} = \mathsf{P}_{\iota_{K^\circ}}(\mathbf{z})$.

Motivated by the above generalization, Moreau had the remarkable vision to prove the following celebrated results:

Theorem 3 (Moreau [1965]) *Let $f \in \Gamma_0$ and $f^*(\mathbf{y}) := \sup_{\mathbf{x}\in\mathcal{H}} \langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})$ its Fenchel conjugate. Then the following are equivalent:*

- $\mathbf{z} = \mathbf{x} + \mathbf{y}, \mathbf{y} \in \partial f(\mathbf{x})$ *(or equivalently $\mathbf{x} \in \partial f^*(\mathbf{y})$, or $f(\mathbf{x}) + f^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle$);*
- $\mathbf{x} = \mathsf{P}_f(\mathbf{z}), \mathbf{y} = \mathsf{P}_{f^*}(\mathbf{z})$.

Theorem 4 (Moreau [1965]) *Let $f \in \Gamma_0$. Then for all $\mathbf{z} \in \mathcal{H}$,*

$$\mathsf{P}_f(\mathbf{z}) + \mathsf{P}_{f^*}(\mathbf{z}) = \mathbf{z} \tag{3}$$

$$\mathsf{M}_f(\mathbf{z}) + \mathsf{M}_{f^*}(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|^2. \tag{4}$$

Before we can say anything about the above results, we need to first define the terms P_f and M_f, which is our goal in the next section.

3 Definitions

For any $f \in \Gamma_0$, we define its corresponding proximity operator and Moreau envelop as follows:

$$\mathsf{P}_f^\eta(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{H}} \tfrac{1}{2\eta}\|\mathbf{x}-\mathbf{z}\|^2 + f(\mathbf{x}) \tag{5}$$

$$\mathsf{M}_f^\eta(\mathbf{z}) = \min_{\mathbf{x}\in\mathcal{H}} \tfrac{1}{2\eta}\|\mathbf{x}-\mathbf{z}\|^2 + f(\mathbf{x}) \tag{6}$$

Intuitively, we try to approximate the point $\mathbf{z}$ with some other point $\mathbf{x}$ under the norm $\|\cdot\|$ and the "penalty" $f(\mathbf{x})$. The positive parameter $\eta > 0$ is introduced as a means to control the approximation. For simplicity, we will consider $\eta = 1$ most of the time, and use the shorthand $\mathsf{P}_f := \mathsf{P}_f^1$ and $\mathsf{M}_f := \mathsf{M}_f^1$.

Some immediate observations:

- The proximity operator P_f is a strict generalization of the projection operator (hence explains our slightly unusual notation for the latter), while the Moreau envelop M_f is a strict generalization of the (squared) distance function. Note that M_f is real-valued (even when f takes ∞) while P_f is $\mathcal{H}$-valued.
- $\mathsf{P}_f(\mathbf{z})$ is *uniquely* determined for any $\mathbf{z} \in \mathcal{H}$ since the squared norm $\|\cdot\|^2$ is strongly convex. Taking derivative we obtain (where Id denotes the identity map)[1]

$$\mathsf{P}_f^\eta = (\mathsf{Id} + \eta\partial f)^{-1}, \tag{7}$$

  namely $\mathbf{x} = \mathsf{P}_f^\eta(\mathbf{z}) \iff \mathbf{z} \in \mathbf{x} + \eta\partial f(\mathbf{x})$, thanks to the uniqueness of the minimizer.

It is possible (and sometimes desirable) to replace the squared norm in the definitions (5) and (6) with some other "distance" function, such as a Bregman divergence.

Example 1 (Huber's function and the shrinkage operator) *Let $f(\mathbf{x}) = \|\mathbf{x}\|$, then*

$$\mathsf{P}_f^\eta(\mathbf{z}) = (1 - \eta/\|\mathbf{z}\|)_+ \cdot \mathbf{z} \tag{8}$$

$$\mathsf{M}_f^\eta(\mathbf{z}) = \begin{cases} \frac{1}{2\eta}\|\mathbf{z}\|^2, & \|z\| \le \eta \\ \|\mathbf{z}\| - \frac{\eta}{2}, & \|z\| \ge \eta \end{cases}. \tag{9}$$

Take $\mathcal{H} = \mathbb{R}$ and $f(x) = |x|$ we obtain the shrinkage operator and the Huber's function, respectively. See Figure 1 for an illustration.

Example 2 (Projection to the 1-norm ball) *By Theorem 4, $\mathsf{P}_{\iota_{\|\cdot\|_1\le 1}} = \mathsf{Id} - \mathsf{P}_{\|\cdot\|_\infty}$, while the latter can be easily computed as follows:*

$$\mathsf{M}_{\|\cdot\|_\infty}(\mathbf{y}) = \min_{t\in\mathbb{R}_+} \min_{|x_i|\le t} \tfrac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2 + t.$$

This is a piecewise quadratic function of the scalar t. We can solve it by careful bookkeeping.

[1] It is a highly nontrivial fact, due to Minty, that the map $\mathsf{Id} + \partial f : \operatorname{dom} f \to \mathcal{H}$ is onto.

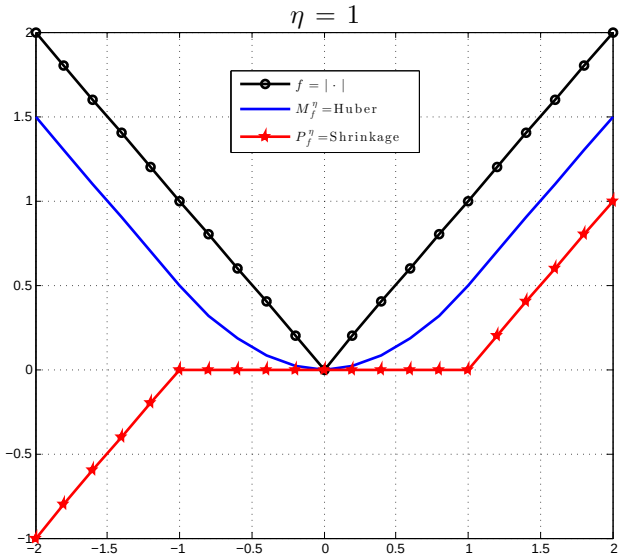

Figure 1: The Moreau envelop and proximity operator of the absolute function.

4 Properties

We are now ready to prove Theorem 4.

Proof (of Theorem 4): Let $\mathbf{x} = \mathsf{P}_f(\mathbf{z}), \mathbf{y} = \mathsf{P}_{f^*}(\mathbf{z})$. Then $\mathbf{z} - \mathbf{y} \in \partial f^*(\mathbf{y}) \iff \mathbf{y} \in \partial f(\mathbf{z} - \mathbf{y}) \iff \mathbf{z} \in \mathbf{z} - \mathbf{y} + \partial(\mathbf{z} - \mathbf{y}) \iff \mathbf{x} = \mathbf{z} - \mathbf{y}$.

For the Moreau envelop, we have

$$\mathsf{M}_f(\mathbf{z}) + \mathsf{M}_{f^*}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{x},\mathbf{y}\in\mathcal{H}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 + \tfrac{1}{2}\|\mathbf{z}-\mathbf{y}\|^2 + f(\mathbf{x}) + f^*(\mathbf{y}) \tag{10}$$

$$\geq \operatorname*{argmin}_{\mathbf{x},\mathbf{y}\in\mathcal{H}} \tfrac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 + \tfrac{1}{2}\|\mathbf{z}-\mathbf{y}\|^2 + \langle \mathbf{x}, \mathbf{y}\rangle \tag{11}$$

$$= \tfrac{1}{2}\|\mathbf{z}\|^2. \tag{12}$$

We can verify that the equality is attained. This completes the proof of Theorem 4.

The proof of Theorem 3 is similar (in fact easier). We have no difficulty in verifying that Theorem 2 is indeed a special case of Theorem 3.

This next result makes intuitive sense, although we shall not prove it here.

Theorem 5 $\mathsf{M}_f^\eta \to f$ *pointwise as* $\eta \to 0$*; moreover, the convergence is actually uniform if* f *is Lipschitz continuous (w.r.t. the norm* $\|\cdot\|$*).*

Theorem 6 (Moreau [1965]) *The Moreau envelop is (Frechét) differentiable, with* $\nabla \mathsf{M}_f = \mathsf{Id} - \mathsf{P}_f = \mathsf{P}_{f^*}$.

Theorem 7 (Moreau [1965]) $\mathsf{P}_f : (\mathcal{H}, \|\cdot\|) \to (\mathcal{H}, \|\cdot\|)$ *is a (firm) non-expansion, namely* 1*-Lipschitz continuous.*

Recall that the projection operator is a non-expansion. Thus we see that the proximity operator is indeed a very natural generalization that retains many useful properties of the projection operator.

Theorem 8 *For any* $\eta > 0$, $\mathbf{x} \in \operatorname{argmin} f \iff \mathbf{x} \in \operatorname{argmin} \mathsf{M}_f^\eta \iff \mathbf{x} = \mathsf{P}_f^\eta(\mathbf{x})$.

Proof: $\mathbf{0} \in \partial f(\mathbf{x}) \iff \mathbf{x} \in \mathbf{x} + \eta \cdot \partial f(\mathbf{x}) \iff \mathbf{x} - \mathsf{P}_f^\eta(\mathbf{x}) = 0 \iff \nabla \mathsf{M}_f^\eta(\mathbf{x}) = 0$. ■

We end this section with an extremely important equivalence between the "regularizer" f and its proximity operator P_f^η. Basically there is no loss of information from f to P_f^η. Thus instead of designing a suitable regularizer, we can directly look for its proximity operator, which, in many cases, underlies the success of our regularizer.

Theorem 9 (Moreau [1965]) *Let $f, g \in \Gamma_0$. Then the following are equivalent:*

- *$f = g + c$ for some constant $c \in \mathbb{R}$;*
- *$\partial f(\mathbf{x}) \subseteq \partial g(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{H}$;*
- *$\mathsf{P}_f^\eta(\mathbf{z}) = \mathsf{P}_g^\eta(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{H}$ and any $\eta > 0$.*

Proof: Assuming $\partial f \subseteq \partial g$, then $(\mathsf{Id} + \eta \partial f)^{-1} \subseteq (\mathsf{Id} + \eta \partial g)^{-1}$. By the uniqueness of the proximity operator, we must have $\mathsf{P}_f^\eta = \mathsf{P}_g^\eta$.

Next assume $\mathsf{P}_f^\eta = \mathsf{P}_g^\eta$. By Theorem 6 we know $\mathsf{M}_f^\eta = \mathsf{M}_g^\eta - c$ for some constant $c \in \mathbb{R}$. Repeatedly conjugating we obtain $f = g + c$. ■

Moreau [1965] also gives a sufficient and necessary characterization of the proximity operator. However, a *convenient* characterization (or even a *convenient* sufficient condition) is yet to be found.

5 Second Motivation

Let us consider solving the problem

$$\min_{\mathbf{x} \in \mathcal{H}} f(\mathbf{x}).$$

However, the objective function f could be non-smooth or ugly.

Idea: Replace f with the smooth envelop function M_f^η, i.e.

$$\left\{ \min_{\mathbf{y} \in \mathcal{H}} \mathsf{M}_f^\eta(\mathbf{y}) \right\} = \left\{ \min_{\mathbf{y} \in \mathcal{H}} \min_{\mathbf{x} \in \mathcal{H}} f(\mathbf{x}) + \tfrac{1}{2\eta} \|\mathbf{x} - \mathbf{y}\|^2 \right\}.$$

Justification: Theorem 8 tells us that the minimizers of f are preserved; Theorem 6 and Theorem 7 together implies that the envelop function M_f^η has $1/\eta$ Lipschitz continuous gradient.

Can alternate between $\mathbf{x}$ and $\mathbf{y}$. This idea, known as the proximal point algorithm [Martinet, 1970, Rockafellar, 1976], leads to valuable insights. Note that nothing prevents us from changing η from step to step.

Chicken-egg problem: But, solving the inner problem $\min_{\mathbf{x} \in \mathcal{H}} f(\mathbf{x}) + \frac{1}{2\eta} \|\mathbf{x} - \mathbf{y}\|^2$ for any fixed $\mathbf{y}$ in general can be as hard as minimizing f directly! The upside is that we gain $1/\eta$-strong convexity which might significantly improve the condition number of our problem. Particularly true if f is some ill-conditioned quadratic function.

6 Composite Problem

The following problem is of considerable importance in machine learning:

$$\min_{\mathbf{x} \in \mathcal{H}} \ell(\mathbf{x}) + f(\mathbf{x}), \tag{13}$$

where ℓ is usually some loss function while f represents some regularizer, such as the 1-norm that promotes sparsity. The regularizer f is desirably nonsmooth, hence computationally harder to deal with.

The proximal gradient algorithm can be derived from the optimality condition $0 \in \nabla \ell(\mathbf{x}) + \partial f(\mathbf{x}) \iff \mathbf{x} - \eta \nabla \ell(\mathbf{x}) \in \mathbf{x} + \eta \partial f(\mathbf{x}) \iff \mathbf{x} = \mathsf{P}_f^\eta(\mathbf{x} - \eta \nabla \ell(\mathbf{x}))$, thus motivating the iteration:

$$\mathbf{y} \leftarrow \mathbf{x} - \eta \cdot \nabla \ell(\mathbf{x}) \tag{14}$$

$$\mathbf{x} \leftarrow \mathsf{P}_f^\eta(\mathbf{y}). \tag{15}$$

Note that the above is true for any $\eta > 0$, although for stability reasons we might have to restrict η. Can prove convergence rates under various assumptions on the loss ℓ. However, the success of this algorithm relies on our ability to compute the proximity operator P_f efficiently (ideally in closed-form).

Goldstein [1964] (two-page!) proposed and proved the convergence of the projected gradient algorithm. Perhaps a little surprisingly, his argument goes through for the proximal gradient algorithm without a single change! This confirms again how natural the proximity operator is as a generalization of the projection operator.

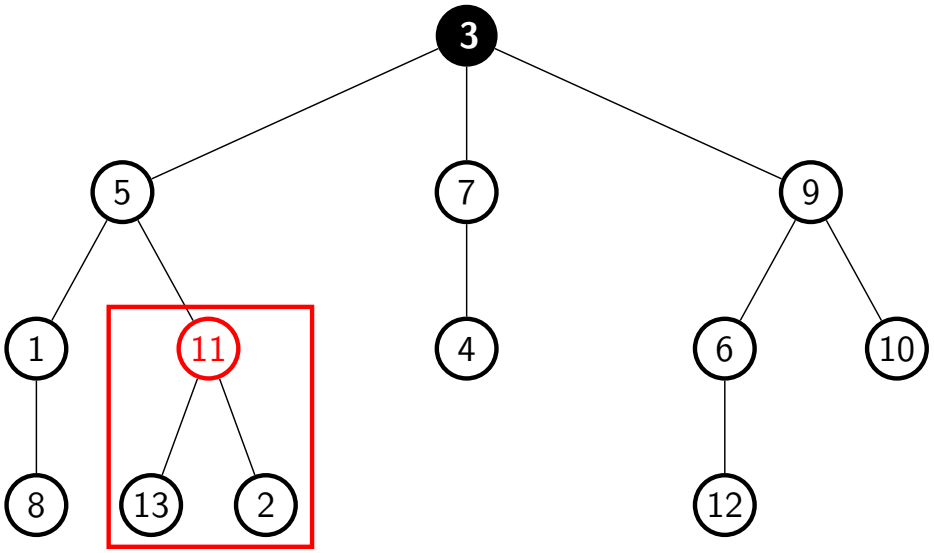

Figure 2: Tree-structured groups are simply rooted subtrees in a rooted tree. The red rectangle denotes the group induced by the subtree rooted at the red node.

7 Smoothing

What if the loss ℓ is nonsmooth, say the Hinge loss? Recycle the idea in Section 5: Replace ℓ with M_ℓ^η. By Theorem 5, if ℓ is Lipschitz continuous (such as the Hinge loss), M_f^η is a uniform approximation of f, hence will not affect the minimization much, provided that we set η properly.

8 Prox-decomposition

What if the regularizer f can be split into the sum of two components, each of whose proximity operators can be easily computed? Can we leverage this fact to come up with an efficient way to compute P_f? This will allow us to *combine* regularizers in a computationally friendly way.

The following result, although being exact and general, has limited use in practice.

Theorem 10 ( ) $\mathsf{P}_{f+g} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id})$.

Proof: Indeed, assuming $\partial(f+g) = \partial f + \partial g$, we have

$$\mathsf{P}_{f+g} = (\mathsf{Id} + \partial(f+g))^{-1} = (\mathsf{Id} + \partial f + \partial g)^{-1} = \left[\frac{(\mathsf{Id} + 2\partial f) + (\mathsf{Id} + 2\partial g)}{2}\right]^{-1}$$
$$= \left[\frac{\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1}}{2}\right]^{-1} = (\mathsf{P}_{2f}^{-1} + \mathsf{P}_{2g}^{-1})^{-1} \circ (2\mathsf{Id}).$$

■

Nevertheless, the next two results are motivating:

Example 3 (Fused LASSO) *Friedman et al. [2007] showed that* $\mathsf{P}_{\|\cdot\|_1 + \|\cdot\|_{\mathsf{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathsf{TV}}}$, *where* $\mathcal{H} = \mathbb{R}^d$, $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$, $\|\mathbf{x}\|_{\mathsf{TV}} = \sum_{i=1}^{d-1} |x_i - x_{i+1}|$.

Example 4 (Tree-structured group LASSO, Jenatton et al. [2011]) *Assuming the groups* $\{\mathsf{g}_i\}$ *form a laminar system* ($\mathsf{g}_i \cap \mathsf{g}_j \in \{\mathsf{g}_i,\ \mathsf{g}_j,\ \emptyset\}$), *then, if appropriately ordered,*

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

where $\|\cdot\|_{\mathsf{g}_i}$ *is the restriction of p-norm,* $p \in \{1, 2, \infty\}$ *to the group* g_i.

See Figure 2 for an illustration.

Therefore, we are motivated to look for the formula

$$\mathsf{P}_{f+g} \overset{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \overset{?}{=} \mathsf{P}_g \circ \mathsf{P}_f. \tag{16}$$

Note that the ordering on the right hand side might have an effect, for the left hand side is invariant to the ordering. Unfortunately, in general, the above decomposition need not hold. Counterexamples can be found in Yu [2013a]. However, we do have a simple, convenient sufficient condition.

Theorem 11 *A sufficient condition for* $\mathsf{P}_{f+g}(\mathbf{z}) = \mathsf{P}_f\big(\mathsf{P}_g(\mathbf{z})\big)$ *for all* $\mathbf{z} \in \mathcal{H}$ *is that*

$$\forall\ \mathbf{y} \in \operatorname{dom} g,\ \ \partial g(\mathsf{P}_f(\mathbf{y})) \supseteq \partial g(\mathbf{y}). \tag{17}$$

Proof: Taking derivative we have

$$\mathsf{P}_{f+g}(\mathbf{z}) - \mathbf{z} + \partial(f+g)(\mathsf{P}_{f+g}(\mathbf{z})) \ni 0 \tag{18}$$

$$\mathsf{P}_g(\mathbf{z}) - \mathbf{z} + \partial g(\mathsf{P}_g(\mathbf{z})) \ni 0 \tag{19}$$

$$\mathsf{P}_f(\mathsf{P}_g(\mathbf{z})) - \mathsf{P}_g(\mathbf{z}) + \partial f(\mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))) \ni 0. \tag{20}$$

Adding the last two equations we obtain

$$\mathsf{P}_f(\mathsf{P}_g(\mathbf{z})) - \mathbf{z} + \partial g(\mathsf{P}_g(\mathbf{z})) + \partial f(\mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))) \ni 0. \tag{21}$$

Now we need only compare (18) and (21). Indeed, let $\mathbf{y} = \mathsf{P}_g(\mathbf{z})$. Then by (21) and the subdifferential rule $\partial(f+g) \supseteq \partial f + \partial g$ we verify that $\mathsf{P}_f(\mathsf{P}_g(\mathbf{z}))$ satisfies the optimality condition (18), hence follows $\mathsf{P}_{f+g}(\mathbf{z}) = \mathsf{P}_f\big(\mathsf{P}_g(\mathbf{z})\big)$ since the proximal map is single-valued. ■

The next result is a reassurance of the impossibility to have the decomposition (16). We omit the proof here.

Theorem 12 (Yu [2013a]) *Fix* $f \in \Gamma_0$. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ *for all* $g \in \Gamma_0$ *iff*

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

Similarly, if we fix $g \in \Gamma_0$*, then* $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ *for all* $f \in \Gamma_0$ *iff* g *is a continuous affine function.*

The first invariant property we consider is scaling-invariance. What kind of convex functions have their subdifferential invariant to (positive) scaling, that is $\partial g(\lambda \cdot \mathbf{w}) = \partial g(\mathbf{w})$ for all $\lambda > 0$? Assuming $\mathbf{0} \in \operatorname{dom} g$ and by simple integration[2]

$$g(t\mathbf{z}) - g(\mathbf{0}) = \int_0^t g'(s\mathbf{z})\,\mathrm{d}s = \int_0^t \langle \mathbf{z}, \partial g(s\mathbf{z})\rangle\,\mathrm{d}s = t \cdot [g(\mathbf{z}) - g(\mathbf{0})],$$

where the last equality follows from the scaling invariance of the subdifferential of g. Therefore, up to some additive constant, g is positively homogeneous (p.h.). On the other hand, if $g \in \Gamma_0$ is p.h. (automatically $0 \in \operatorname{dom} g$), then from definition we verify that ∂g is scaling-invariant. Therefore, under the scaling-invariance assumption, g consists of all p.h. functions in Γ_0, up to some additive constant. Consequently, the requirement on f is to have its proximal map $\mathsf{P}_f(\mathbf{z}) = \lambda_{\mathbf{z}} \cdot \mathbf{z}$ for some $\lambda_{\mathbf{z}} \in [0,1]$ that may depend on $\mathbf{z}$ as well[3]. The next theorem completely characterizes such functions.

Theorem 13 (Yu [2013a]) *Let* $f \in \Gamma_0$. *Consider the statements*

1. $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$*;*
2. *For all perpendicular* $\mathbf{x} \perp \mathbf{y} \implies f(\mathbf{x}+\mathbf{y}) \geq f(\mathbf{y})$*;*

[2]Here $g'(s\mathbf{z})$, as a function of the scalar s, denotes its right derivative, or, thanks to the convexity of g and the "robustness" of integration, any other sensible selection of the subdifferential.

[3]Note that $\lambda_{\mathbf{z}} \leq 1$ is necessary since any proximal map is nonexpansive.

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

4. $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all p.h. (up to some additive constant) functions* $\kappa \in \Gamma_0$.

Then we have 1) $\implies$ *2)* $\iff$ *3)* $\iff$ *4). Moreover, when* $\dim(\mathcal{H}) \geq 2$, *2)* $\implies$ *1) as well, in which case* $\mathsf{P}_f(\mathbf{z}) = \mathsf{P}_h(\|\mathbf{z}\|)/\|\mathbf{z}\| \cdot \mathbf{z}$ *(where we interpret* $0/0 = 0$*).*

Quite unexpectedly, the first two conditions in Theorem 13 characterizes the representer theorem for a closed convex regularizer [Dinuzzo and Schölkopf, 2012]. Therefore we deduce from Theorem 13 that the representer theorem is equivalent to a computational requirement (the last two items above)!

Example 5 (Elastic net [Zou and Hastie, 2005]) *As usual, denote* $\mathsf{q} = \frac{1}{2}\|\cdot\|^2$. *In many applications, in addition to the regularizer* κ *(usually a p.h. convex function), one adds the squared 2-norm regularizer* $\lambda\mathsf{q}$ *for stability, grouping effect, strong convexity, etc. This incurs no computational cost in the sense of computing the proximity operator: We easily compute that* $\mathsf{P}_{\lambda\mathsf{q}} = \frac{1}{\lambda+1}\mathsf{Id}$. *By Theorem 13, for any p.h. convex function* κ, $\mathsf{P}_{\kappa+\lambda\mathsf{q}} = \frac{1}{\lambda+1}\mathsf{P}_\kappa$, *whence it is also clear that adding an extra squared 2-norm regularizer tends to double "shrink" the solution. In particular, let* $\mathcal{H} = \mathbb{R}^d$ *and take* κ *to be the 1-norm, we recover the proximal map for the elastic-net.*

Instead of restricting to each ray, we now generalize our result to cones. Specifically, consider the gauge, that is, a p.h. convex function

$$\kappa(\mathbf{x}) = \max_{j \in J} \langle \mathbf{a}_j, \mathbf{x} \rangle , \tag{22}$$

where J is a finite index set and each $\mathbf{a}_j \in \mathcal{H}$. Such (polyhedral) gauge functions have become extremely important in machine learning. Define the polyhedral cones[4]

$$K_j = \{\mathbf{x} \in \mathcal{H} : \langle \mathbf{a}_j, \mathbf{x} \rangle = \kappa(\mathbf{x})\}. \tag{23}$$

Assume $K_j \neq \emptyset$ for each j (otherwise delete j from J). The sufficient condition (17), with $g = \kappa$, becomes $\partial\kappa(\mathsf{P}_f(\mathbf{y})) \supseteq \partial\kappa(\mathbf{y})$. Since $\partial\kappa(\mathbf{x}) = \{\mathbf{a}_j : j \in J, \mathbf{x} \in K_j\}$, $\mathbf{a}_j \in \partial\kappa(\mathbf{y}) \iff \mathbf{y} \in K_j$, hence $\mathbf{a}_j \in \partial\kappa(\mathsf{P}_f(\mathbf{y})) \iff \mathsf{P}_f(\mathbf{y}) \in K_j$. In other words, we simplify the sufficient condition (17) as

$$\forall j \in J,\ \mathbf{y} \in K_j \implies \mathsf{P}_f(\mathbf{y}) \in K_j. \tag{24}$$

That is, each cone K_j is "fixed" under the proximity operator of f. Instead of completely characterizing f under (24), we show that in its current form, (24) is already very interesting.

Recall that a set function $\mu : 2^{[d]} \to \mathbb{R}$ is submodular if for all $A, B \subseteq [d]$, we have

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

The Choquet integral (a.k.a. the Lovász extension) of μ is defined as the function $f : \mathbb{R}^d \to \mathbb{R}$ with

$$\mathbf{w} \mapsto \int_0^\infty \mu([\![\mathbf{w} \geq t]\!])\,\mathrm{d}t + \int_{-\infty}^0 \Big(\mu([\![\mathbf{w} \geq t]\!]) - \mu([d])\Big)\,\mathrm{d}t. \tag{25}$$

It can be proved that f is convex iff μ is submodular. If μ is an additive measure, then the Choquet integral is simply the Lebesgue integral (if we treat each vector $\mathbf{w}$ as a function from $[d]$ to $\mathbb{R}$). When $\mathbf{w} \geq 0$, (25) is the familiar "integrating the tail" trick to compute the expectation.

Theorem 14 *Let* g *be the Choquet integral of some submodular function, then* $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ *for all permutation invariant function* $f \in \Gamma_0$.

Proof: (Sketch): We show that the subdifferential of any convex Choquet integral is determined by the relative ordering of the input vector $\mathbf{x}$, while $\mathsf{P}_f(\mathbf{z})$ has the same relative ordering as $\mathbf{z}$. Therefore we verify (24). ■

There exists simple, convenient characterizations of the Choquet integral. For instance, all total variation norms are convex Choquet integrals, hence Example 3 is a special case of the above result. Of course it is possible to derive many more cases from Theorem 14.

[4] A set is polyhedral if it is the intersection of *finitely* many half spaces. Polyhedral sets are closed convex.

9 Proximal Averaging

When the sufficient condition in Theorem 11 fails, we can resort to the following simple *linearization* idea:

$$\mathsf{P}^{\eta}_{\sum_i f_i} \approx \sum_i \mathsf{P}^{\eta}_{f_i}. \tag{26}$$

It can be shown that there exists a function, called the proximal average, whose proximity operator is exactly the right hand side above. Therefore as long as we can show that the proximal average is close to our original function $\sum_i f_i$, we will be fine with the above approximation. Details can be found in Yu [2013b].

References

Francesco Dinuzzo and Bernhard Schölkopf. The representer theorem for Hilbert spaces: a necessary and sufficient condition. In *Advances in Neural Information Processing Systems*, 2012.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

A. A. Goldstein. Convex programming in hilbert space. *Bulletin of the American Mathematical Society*, 70 (5):709–710, 1964.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Revue Française d'Informatique et de Recherche Opérationnelle, Série Rouge*, 4(3):154–158, 1970.

Jean J. Moreau. Proximité et dualtité dans un espace Hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Ralph Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control*, 14(5):877–898, 1976.

Yaoliang Yu. On decomposing the proximal map. In *Advances in Neural Information Processing Systems*, 2013a.

Yaoliang Yu. Better approximation and faster algorithm using the proximal average. In *Advances in Neural Information Processing Systems*, 2013b.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.