The proximal mapping

Acknowledgement: slides are based on Prof. Lieven Vandenberghes and Prof. Zaiwen Wen.

Outline

1. Closed function
2. Conjugate function
3. Proximal mapping

Closed set

A set $\mathcal{C}$ is closed if it contains its boundary:

$$x^k \in \mathcal{C}, \quad x^k \to \bar{x} \quad \Rightarrow \quad \bar{x} \in \mathcal{C}$$

Operations that preserve closedness

- the intersection of closed sets is closed
- the union of a finite number of closed sets is closed
- inverse under linear mapping: $\{x | Ax \in \mathcal{C}\}$ is closed if $\mathcal{C}$ is closed

Image under linear mapping

The image of a closed set under a linear mapping is not necessarily closed
example ($\mathcal{C}$ is closed, $A\mathcal{C} = \{Ax | x \in \mathcal{C}\}$ is open:)

$$\mathcal{C} = \{(x_1, x_2) \in \mathbb{R}^2_+ | x_1 x_2 \geq 1\}, \quad A = [1, 0], A\mathcal{C} = \mathbb{R}_{++}$$

sufficient condition: $A\mathcal{C}$ is closed if

- $\mathcal{C}$ is closed and convex
- and $\mathcal{C}$ does not have a recession direction in the null space of A, i.e.

$$Ay = 0, \hat{x} \in \mathcal{C}, \hat{x} + \alpha y \in \mathcal{C}, \forall \alpha > 0 \quad \Rightarrow y = 0.$$

in particular, this holds for any A if $\mathcal{C}$ is bounded.

Closed function

definition: a function is closed if its epigraph is a closed set or if all its sublevel set is a closed set.

- If f is continuous and $\mathrm{dom} f$ is closed, then f is closed
- If f is continuous and $\mathrm{dom} f$ is open, then f is closed iff it converges to ∞ along every sequence converging to a boundary point of $\mathrm{dom} f$

examples

- $f(x) = x \log x$ with $\mathrm{dom} f = \mathbb{R}_+$ and $f(0) = 0$
- indicator function of a closed set.

not closed

- $f(x) = x \log x$ with $\mathrm{dom} f = \mathbb{R}_{++}$ or $\mathrm{dom} f = \mathbb{R}_+$ and $f(0) = 1$
- indicator function of a set $\mathcal{C}$ if $\mathcal{C}$ is not closed

Properties

sublevel sets: f is closed iff all its subsevel sets are closed

minimum: if f is closed with bounded sublevel sets then it has a minimizer.

Theorem (Weierstrass)

Suppose that the set $\mathcal{D} \subset \mathcal{E}$ (a finite dimensional vector space over $\mathbb{R}^n$) is the nonempty and closed, ant that all sublevel sets of a continuous function $f : \mathcal{D} \mapsto \mathbb{R}$ are bounded. Then f has a global minimizer.

Operation that preserves closedness on convex functions

- $f + g$ is closed if f and g are closed (and $\mathrm{dom} f \cap \mathrm{dom} g \neq \emptyset$)
- $f(Ax + b)$ is closed if f is closed
- $\sup_\alpha f_\alpha(x)$ is closed if each function f_α is closed.

Conjugate functions: recall

the **conjugate** of a function f is

$$f^*(y) = \sup_{x \in \mathrm{dom} f} (y^\top x - f(x))$$

f^* is closed and convex even if f is not

Fenchel's inequality

$$f(x) + f^*(y) \geq x^\top y, \forall x, y$$

(extends inequality $x^\top x/2 + y^\top y/2 \geq x^\top y$ to non-quadratic convex f)

Quadratic function

$$f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$$

strictly convex case $A \succ 0$

$$f^*(y) = \frac{1}{2}(y-b)^\top A^{-1}(y-b) - c$$

general convex case $A \succeq 0$

$$f^*(y) = \frac{1}{2}(y-b)^\top A^\dagger(y-b) - c, \quad \mathrm{dom} f^* = \mathrm{Range}(A) + b$$

Negative entropy and negative logarithm

Negative entropy

$$f(x) = \sum_{i=1}^{n} x_i \log x_i, \quad f^*(y) = \sum_{i=1}^{n} \exp(y_i - 1)$$

Negative logarithm

$$f(x) = -\sum_{i=1}^{n} \log x_i, \quad f^*(y) = -\sum_{i=1}^{n} \log(-y_i) - n$$

Matrix logarithm

$$f(X) = -\log\det(X) \quad (\mathrm{dom} f = \mathbb{S}^n_{++}),\ f^*(Y) = -\log\det(-Y) - n$$

Indicator function

The indicator function of convex set $\mathcal{C}$: conjugate is support function of $\mathcal{C}$

$$f(x) = \begin{cases} 0, & x \in \mathcal{C} \\ +\infty, & x \notin \mathcal{C} \end{cases}, \quad f^*(y) = \sup\{y^\top x | x \in \mathcal{C}\}.$$

Norm: conjugate is indicator of unit dual norm ball

$$f(x) = \|x\|, \quad f^*(y) = \begin{cases} 0, & \|y\|_* \leq 1 \\ +\infty, & \|y\|_* > 1 \end{cases}$$

Recall the definition of dual norm: $\|y\|_* = \sup\{x^\top y | \|x\| \leq 1\}$.

The second conjugate

$$f^{**}(x) = \sup_{y \in \mathrm{dom} f^*} (x^\top y - f^*(y))$$

- f^{**} is closed and convex
- From Fenchel's inequality $x^\top y - f^*(y) \leq f(x)$ for all y and x:

  $$f^{**}(x) \leq f(x), \quad \forall x,$$

  equivalently, $\mathrm{epi} f \subseteq \mathrm{epi} f^{**}$ for any f
- if f is closed and convex, then $f^{**} = f$.

Conjugates and subgradients

if f is closed and convex, then

$$y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y) \Leftrightarrow x^\top y = f(x) + f^*(y)$$

Proof:if $y \in \partial f(x)$, then $f^*(y) = \sup_u (y^\top u - f(u)) = y^\top x - f(x)$

$$\begin{aligned} f^*(v) &= \sup_u (v^\top u - f(u)) \geq v^\top x - f(x) \\ &= x^\top (v - y) - f(x) + y^\top x \\ &= f^*(y) + x^\top (v - y) \end{aligned}$$

for all v; therefore $x \in \partial f^*(y)$.

Reverse implication $x \in \partial f^*(y) \Rightarrow y \in \partial f(x)$ follows from $f^{**} = f$

Some calculus rules

Separable sum

$$f(x_1, x_2) = g(x_1) + h(x_2), \quad f^*(y_1, y_2) = g^*(y_1) + h^*(y_2)$$

Scalar multiplication ($\alpha > 0$)

$$f(x) = \alpha g(x), \quad f^*(y) = \alpha g^*(y/\alpha)$$

addition to affine function

$$f(x) = g(x) + a^\top x + b \quad f^*(y) = g^*(y - a) - b$$

infimal convolution

$$f(x) = \inf_{u+v=x}(g(u) + h(v)) \quad f^*(y) = g^*(y) + h^*(y)$$

Proximal mapping

Definition: the proximal mapping of a closed convex function f is

$$\operatorname{prox}_f(x) = \underset{u}{\operatorname{argmin}} \left(f(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

Existence and uniqueness: we minimize a closed and strongly convex function

$$g(u) = f(u) + \frac{1}{2}\|u - x\|_2^2$$

- minimizer exists because g is closed with bounded sublevel sets
- minimizer is unique because g is strictly convex

Subgradient characterization (from page 4.7):

$$u = \operatorname{prox}_f(x) \qquad \Longleftrightarrow \qquad x - u \in \partial f(u)$$

Examples

Quadratic function ($A \succeq 0$)

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c, \qquad \mathrm{prox}_{tf}(x) = (I + tA)^{-1}(x - tb)$$

Euclidean norm: $f(x) = \|x\|_2$

$$\mathrm{prox}_{tf}(x) = \begin{cases} (1 - t/\|x\|_2)x & \|x\|_2 \geq t \\ 0 & \text{otherwise} \end{cases}$$

Logarithmic barrier

$$f(x) = -\sum_{i=1}^{n} \log x_i, \qquad \mathrm{prox}_{tf}(x)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}, \quad i = 1, \ldots, n$$

Simple calculus rules

Separable sum

$$f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = g(x) + h(y), \qquad \mathrm{prox}_f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \mathrm{prox}_g(x) \\ \mathrm{prox}_h(y) \end{bmatrix}$$

Scaling and translation of argument: for scalar $a \neq 0$,

$$f(x) = g(ax + b), \qquad \mathrm{prox}_f(x) = \frac{1}{a}\left(\mathrm{prox}_{a^2 g}(ax + b) - b\right)$$

"Right" scalar multiplication: with $\lambda > 0$,

$$f(x) = \lambda g(x/\lambda), \qquad \mathrm{prox}_f(x) = \lambda \mathrm{prox}_{\lambda^{-1} g}(x/\lambda)$$

Addition to linear or quadratic function

Linear function

$$f(x) = g(x) + a^T x, \qquad \mathrm{prox}_f(x) = \mathrm{prox}_g(x - a)$$

Quadratic function: with $\mu > 0$

$$f(x) = g(x) + \frac{\mu}{2}\|x - a\|_2^2, \qquad \mathrm{prox}_f(x) = \mathrm{prox}_{\theta g}(\theta x + (1 - \theta)a),$$

where $\theta = 1/(1 + \mu)$

Moreau decomposition

$$x = \mathrm{prox}_f(x) + \mathrm{prox}_{f^*}(x) \quad \text{for all } x$$

- follows from properties of conjugates and subgradients:

$$\begin{aligned} u = \mathrm{prox}_f(x) \quad &\Longleftrightarrow \quad x - u \in \partial f(u) \\ &\Longleftrightarrow \quad u \in \partial f^*(x - u) \\ &\Longleftrightarrow \quad x - u = \mathrm{prox}_{f^*}(x) \end{aligned}$$

- generalizes decomposition by orthogonal projection on subspaces:

$$x = P_L(x) + P_{L^\perp}(x)$$

if L is a subspace, $L^\perp$ its orthogonal complement

(this is the Moreau decomposition with $f = \delta_L$, $f^* = \delta_{L^\perp}$)

Extended Moreau decomposition

for $\lambda > 0$,

$$x = \mathrm{prox}_{\lambda f}(x) + \lambda \mathrm{prox}_{\lambda^{-1} f^*}(x/\lambda) \quad \text{for all } x$$

Proof: apply Moreau decomposition to λf

$$\begin{aligned} x &= \mathrm{prox}_{\lambda f}(x) + \mathrm{prox}_{(\lambda f)^*}(x) \\ &= \mathrm{prox}_{\lambda f}(x) + \lambda \mathrm{prox}_{\lambda^{-1} f^*}(x/\lambda) \end{aligned}$$

second line uses $(\lambda f)^*(y) = \lambda f^*(y/\lambda)$ and expression on page 6.4

Composition with affine mapping

$$f(x) = g(Ax + b)$$

- for general A, prox-operator of f does not follow easily from prox-operator of g
- however, if $AA^T = (1/\alpha)I$, then

$$\begin{aligned} \mathrm{prox}_f(x) &= (I - \alpha A^T A)x + \alpha A^T(\mathrm{prox}_{\alpha^{-1}g}(Ax + b) - b) \\ &= x - \alpha A^T(Ax + b - \mathrm{prox}_{\alpha^{-1}g}(Ax + b)) \end{aligned}$$

Example: $f(x_1, \ldots, x_m) = g(x_1 + x_2 + \cdots + x_m)$

- write as $f(x) = g(Ax)$ with $A = \begin{bmatrix} I & I & \cdots & I \end{bmatrix}$
- since $AA^T = mI$, we get

$$\mathrm{prox}_f(x_1, \ldots, x_m)_i = x_i - \frac{1}{m}\sum_{j=1}^{m} x_j + \frac{1}{m}\mathrm{prox}_{mg}(\sum_{j=1}^{m} x_j), \quad i = 1, \ldots, m$$

Proof: $u = \operatorname{prox}_f(x)$ is the solution of the optimization problem

$$\begin{array}{ll} \text{minimize} & g(y) + \frac{1}{2}\|u - x\|_2^2 \\ \text{subject to} & Au + b = y \end{array}$$

with variables u, y

- eliminate u using the expression

$$\begin{aligned} u &= x + A^T(AA^T)^{-1}(y - b - Ax) \\ &= (I - \alpha A^T A)x + \alpha A^T(y - b) \qquad \text{(since } AA^T = (1/\alpha)I\text{)} \end{aligned}$$

- optimal y is minimizer of

$$g(y) + \frac{\alpha^2}{2}\|A^T(y - b - Ax)\|_2^2 = g(y) + \frac{\alpha}{2}\|y - b - Ax\|_2^2$$

  solution is $y = \operatorname{prox}_{\alpha^{-1}g}(Ax + b)$

Outline

- conjugate functions
- proximal mapping
- **projections**
- support functions, norms, distances

Projection on affine sets

Hyperplane: $C = \{x \mid \ a^T x = b\}$ (with $a \neq 0$)

$$P_C(x) = x + \frac{b - a^T x}{\|a\|_2^2}\, a$$

Affine set: $C = \{x \mid Ax = b\}$ (with $A \in \mathbf{R}^{p \times n}$ and $\mathbf{rank}(A) = p$)

$$P_C(x) = x + A^T(AA^T)^{-1}(b - Ax)$$

inexpensive if $p \ll n$, or $AA^T = I$, ...

Projection on simple polyhedral sets

Halfspace: $C = \{x \mid a^T x \leq b\}$ (with $a \neq 0$)

$$P_C(x) = x + \frac{b - a^T x}{\|a\|_2^2} a \quad \text{if } a^T x > b, \qquad P_C(x) = x \quad \text{if } a^T x \leq b$$

Rectangle: $C = [l, u] = \{x \in \mathbf{R}^n \mid l \leq x \leq u\}$

$$P_C(x)_k = \begin{cases} l_k & x_k \leq l_k \\ x_k & l_k \leq x_k \leq u_k \\ u_k & x_k \geq u_k \end{cases}$$

Nonnegative orthant: $C = \mathbf{R}^n_+$

$$P_C(x) = x_+ = (\max\{0, x_1\}, \max\{0, x_2\}, \ldots, \max\{0, x_n\})$$

Projection on simple polyhedral sets

Probability simplex: $C = \{x \mid \mathbf{1}^T x = 1, x \geq 0\}$

$$P_C(x) = (x - \lambda \mathbf{1})_+$$

where λ is the solution of the equation

$$\mathbf{1}^T (x - \lambda \mathbf{1})_+ = \sum_{i=1}^{n} \max\{0, x_k - \lambda\} = 1$$

Intersection of hyperplane and rectangle: $C = \{x \mid a^T x = b,\ l \leq x \leq u\}$

$$P_C(x) = P_{[l,u]}(x - \lambda a)$$

where λ is the solution of the equation

$$a^T P_{[l,u]}(x - \lambda a) = b$$

Proof (probability simplex): projection $y = P_C(x)$ solves the optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y - x\|_2^2 + \delta_{\mathbf{R}^n_+}(y) \\ \text{subject to} & \mathbf{1}^T y = 1 \end{array}$$

optimality conditions are:

- y minimizes the Lagrangian

$$\begin{aligned} & \frac{1}{2}\|y - x\|_2^2 + \delta_{\mathbf{R}^n_+}(y) + \lambda(\mathbf{1}^T y - 1) \\ = \quad & \sum_{k=1}^{n} \left(\frac{1}{2}(y_k - x_k)^2 + \delta_{\mathbf{R}_+}(y_k) + \lambda y_k \right) - \lambda \end{aligned}$$

  this is a separable function with minimizer $y_k = (x_k - \lambda)_+$ for $k = 1, \ldots, n$

- primal feasibility: requires

$$\sum_{k=1}^{n} y_i = \sum_{k=1}^{n} (x_k - \lambda)_+ = 1$$

Proof (rectangle and hyperplane): $y = P_C(x)$ solves optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y - x\|_2^2 + \delta_{[l,u]}(y) \\ \text{subject to} & a^T y = b \end{array}$$

optimality conditions are:

- y minimizes the Lagrangian

$$\begin{aligned} & \frac{1}{2}\|y - x\|_2^2 + \delta_{[l,u]}(y) + \lambda(a^T y - b) \\ &= \sum_{k=1}^{n} \left(\frac{1}{2}(y_k - x_k)^2 + \delta_{[l_k,u_k]}(y_k) + \lambda a_k y_k \right) - \lambda b \end{aligned}$$

  the minimizer is $y_k = P_{[l_k,u_k]}(x_k - \lambda a_k)$ for $k = 1, \ldots, n$

- primal feasibility: requires

$$a^T y = \sum_{k=1}^{n} a_k P_{[l_k,u_k]}(x_k - \lambda a_k) = b$$

Projection on norm balls

Euclidean ball: $C = \{x \mid \|x\|_2 \leq 1\}$

$$P_C(x) = \frac{1}{\|x\|_2} x \quad \text{if } \|x\|_2 > 1, \qquad P_C(x) = x \quad \text{if } \|x\|_2 \leq 1$$

1-norm ball: $C = \{x \mid \|x\|_1 \leq 1\}$

projection is $P_C(x) = x$ if $\|x\|_1 \leq 1$; otherwise

$$P_C(x)_k = \operatorname{sign}(x_k) \max\{|x_k| - \lambda, 0\} = \begin{cases} x_k - \lambda & x_k > \lambda \\ 0 & -\lambda \leq x_k \leq \lambda \\ x_k + \lambda & x_k < -\lambda \end{cases}$$

where λ is the solution of the equation

$$\sum_{k=1}^{n} \max\{|x_k| - \lambda, 0\} = 1$$

Proof (1-norm): projection $y = P_C(x)$ solves the optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y - x\|_2^2 \\ \text{subject to} & \|y\|_1 \leq 1 \end{array}$$

optimality conditions are:

- y minimizes the Lagrangian

$$\frac{1}{2}\|y - x\|_2^2 + \lambda(\|y\|_1 - \lambda) = \sum_{k=1}^{n} \left(\frac{1}{2}(y_k - x_k)^2 + \lambda|y_k| \right) - \lambda$$

  the minimizer y is obrained by componentwise soft-thresholding:

$$y_k = \operatorname{sign}(x_k) \max\{|x_k| - \lambda, 0\}, \quad k = 1, \ldots, n$$

- primal, dual feasibility and complementary slackness:

$$\lambda = 0, \ \ \|y\|_1 = \|x\|_1 \leq 1 \qquad \text{or} \qquad \lambda > 0, \ \ \|y\|_1 = \sum_{k=1}^{n} \max\{|x_k| - \lambda, 0\} = 1$$

Projection on simple cones

Second order cone: $C = \{(x,t) \in \mathbf{R}^{n \times 1} \mid \|x\|_2 \le t\}$

$$P_C(x,t) = (x,t) \quad \text{if } \|x\|_2 \le t, \qquad P_C(x,t) = (0,0) \quad \text{if } \|x\|_2 \le -t$$

and

$$P_C(x,t) = \frac{t + \|x\|_2}{2\|x\|_2} \begin{bmatrix} x \\ \|x\|_2 \end{bmatrix} \quad \text{if } \|x\|_2 > |t|$$

Positive semidefinite cone: $C = \mathbf{S}_+^n$

$$P_C(X) = \sum_{i=1}^{n} \max\{0, \lambda_i\} q_i q_i^T$$

if $X = \sum_{i=1}^{n} \lambda_i q_i q_i^T$ is the eigenvalue decomposition of X

Outline

- conjugate functions
- proximal mapping
- projections
- **support functions, norms, distances**

Support function

- conjugate of support function of closed convex set is indicator function

$$f(x) = \sup_{y \in C} x^T y, \qquad f^*(y) = \delta_C(y)$$

- prox-operator of support function follows from Moreau decomposition

$$\begin{aligned} \mathrm{prox}_{tf}(x) &= x - t\,\mathrm{prox}_{t^{-1}f^*}(x/t) \\ &= x - tP_C(x/t) \end{aligned}$$

Example: $f(x)$ is sum of largest r components of x

$$f(x) = x_{[1]} + \cdots + x_{[r]} = \delta_C^*(x), \qquad C = \{y \mid 0 \preceq y \preceq \mathbf{1}, \mathbf{1}^T y = r\}$$

prox-operator of f is easily evaluated via projection on C (page 6.12)

Norms

- conjugate of norm is indicator function of dual norm ball:

$$f(x) = \|x\|, \qquad f^*(y) = \delta_B(y) \quad \text{with } B = \{y \mid \|y\|_* \le 1\}$$

- prox-operator of norm follows from Moreau decomposition

$$\begin{aligned} \mathrm{prox}_{tf}(x) &= x - t\,\mathrm{prox}_{t^{-1}f^*}(x/t) \\ &= x - tP_B(x/t) \\ &= x - P_{tB}(x) \end{aligned}$$

- gives $\mathrm{prox}_{t\|\cdot\|}$ when projection on $tB = \{x \mid \|x\|_* \le t\}$ is cheap

Examples: for $\|\cdot\|_1$, $\|\cdot\|_2$, get expressions on pages 4.2 and 6.3

Distance to a point

Distance (in general norm)

$$f(x) = \|x - a\|$$

Prox-operator: from page 6.4, with $g(x) = \|x\|$

$$\begin{aligned} \mathrm{prox}_{tf}(x) &= a + \mathrm{prox}_{tg}(x - a) \\ &= a + x - a - tP_B\left(\frac{x - a}{t}\right) \\ &= x - P_{tB}(x - a) \end{aligned}$$

B is the unit ball for the dual norm $\|\cdot\|_*$

Euclidean distance to a set

Euclidean distance (to a closed convex set C)

$$d(x) = \inf_{y \in C} \|x - y\|_2$$

Prox-operator of distance

$$\mathrm{prox}_{td}(x) = \begin{cases} x + \dfrac{t}{d(x)}(P_C(x) - x) & d(x) \geq t \\ P_C(x) & \text{otherwise} \end{cases}$$

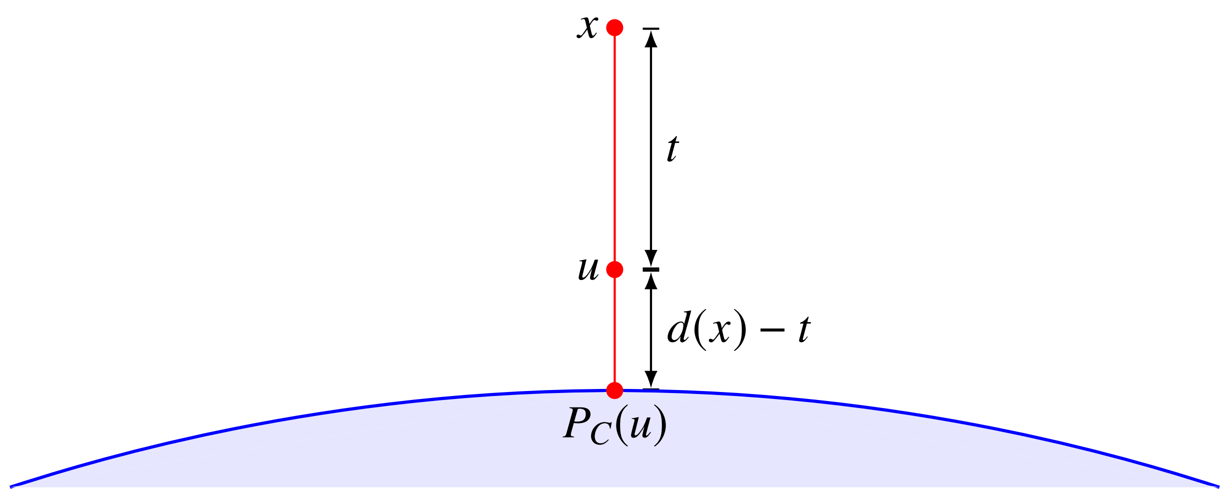
Prox-operator of squared distance: $f(x) = d(x)^2/2$

$$\mathrm{prox}_{tf}(x) = \frac{1}{1+t}x + \frac{t}{1+t}P_C(x)$$

Proof (expression for $\mathrm{prox}_{td}(x)$):

- if $u = \mathrm{prox}_{td}(x) \notin C$, then from page 6.2 and subgradient for d (page 2.20)

$$x - u = \frac{t}{d(u)}(u - P_C(u))$$



- if $\mathrm{prox}_{td}(x) \in C$ then the minimizer of

$$d(u) + \frac{1}{2t}\|u - x\|_2^2$$

satisfies $d(u) = 0$ and must be the projection $P_C(x)$

Proof (expression for $\text{prox}_{tf}(x)$ when $f(x) = d(x)^2/2$):

$$\begin{aligned}
\text{prox}_{tf}(x) &= \underset{u}{\text{argmin}} \left(\frac{1}{2} d(u)^2 + \frac{1}{2t} \|u - x\|_2^2 \right) \\
&= \underset{u}{\text{argmin}} \inf_{v \in C} \left(\frac{1}{2} \|u - v\|_2^2 + \frac{1}{2t} \|u - x\|_2^2 \right)
\end{aligned}$$

- optimal u as a function of v is

$$u = \frac{t}{t+1} v + \frac{1}{t+1} x$$

- optimal v minimizes

$$\frac{1}{2} \left\| \frac{t}{t+1} v + \frac{1}{t+1} x - v \right\|_2^2 + \frac{1}{2t} \left\| \frac{t}{t+1} v + \frac{1}{t+1} x - x \right\|_2^2 = \frac{t}{2(1+t)} \|v - x\|_2^2$$

over C, *i.e.*, $v = P_C(x)$

References

- A. Beck, *First-Order Methods in Optimization* (2017), chapter 6.
- P. L. Combettes and J.-Ch. Pesquet, *Proximal splitting methods in signal processsing*, in: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (2011).
- N. Parikh and S. Boyd, *Proximal algorithms* (2013).