A functional and convex analysis cheat sheet

Guillaume Wang

guillaume.wang@inf.ethz.ch

May 2021

Among the topics that I've been interested in during the last few months, many required some knowledge of convex analysis: regularization in linear models, primal-dual views on optimization, representer theorems in reproducing kernel Banach spaces... Moreover the finite-dimensional setting doesn't suffice, a more abstract point of view is necessary or at least useful; namely Banach spaces seem to be the appropriate level of abstraction for those topics.

In this document I compile some relevant functional and convex analysis background, in the form of a cheat sheet. It is not at all meant to be exhaustive, I only included basic facts and tricks that I found interesting. I may add to it in the future.

1 Functional analysis (Banach duality) cheat sheet

Beyond finite dimension, Banach spaces are a simple and natural level of abstraction for discussing convex analysis. This section is mostly extracted from the appendix of my Master's thesis.

Definition 1.1 (Banach space)**.** A metric space (E, d) is called *complete* if all Cauchy sequences $(u_n)_n \in E^{\mathbb{N}}$ converge in E.

A *Banach space* $(E, \|\cdot\|_E)$ is a vector space equipped with a norm for which it is a complete space.

A *Hilbert space* $(H, \langle\cdot,\cdot\rangle_H)$ is a vector space equipped with an inner product that is complete for the induced norm $\|x\|_H^2 = \langle x, x\rangle_H$.

The unit ball of a normed space $(E, \|\cdot\|_E)$ is denoted $B^{(E)} = B_{0,1}^{(E)} := \{x \in E;\ \ \|x\|_E \leq 1\}$.

A continuous linear mapping $T : E \to F$ between Banach spaces is called a *bounded operator*, and its *operator norm* is the finite quantity $|||T||| = |||T|||_{E\to F} := \sup_{\|x\|_E \leq 1} \|Tx\|_F$. The set of bounded operators from E to F equipped with the operator norm $(\mathcal{L}_b(E, F), |||\cdot|||)$ is itself a Banach space.

A bounded operator $T : E \to F$ is called *compact* if it sends the unit ball into a relatively compact set, i.e $T(B^{(E)})$ is a relatively compact set of F, i.e $\overline{T(B^{(E)})}$ is compact where $\overline{\cdot}$ denotes closure w.r.t the norm of F.

1.1 Duality in Banach spaces

Definition 1.2 (Dual space)**.** The *(topological) dual* of a Banach space E is the space of bounded linear forms $E' = \mathcal{L}_b(E, \mathbb{R})$. It is equipped with the norm $\|X\|_{E'} := \sup_{\|x\|_E \leq 1} |X(x)|$. E' is itself a Banach space.

The *duality bracket* of E is the bilinear operator $\langle\cdot,\cdot\rangle_E : E \times E' \to \mathbb{R}$ defined by $\langle x, X\rangle_E = X(x)$.

The *bidual* of E is the space $E'' = (E')'$. E can be embedded into E'' by $x \mapsto x''$, where x'' is defined by: $\forall X \in E',\ \langle X, x'' \rangle_{E'} = \langle x, X \rangle_E = X(x)$. It is not hard to show (using existence of norming functionals, see below) that this embedding is isometric i.e $\|x\|_{(E')'} = \|x\|_E$.

E is called a *reflexive* Banach space if the converse holds, i.e if any element of the bidual can also be seen as an element of the primal, i.e if $E'' \simeq E$.

In this section, elements of the primal space will typically be denoted by lowercase letters e.g $x \in E, y \in F$, and elements of the dual by uppercase letters e.g $X \in E', Y \in F'$.

Remark 1.1. The duality bracket is very similar to the physicists' bra-ket notation; except that here the primal is on the left and the dual is on the right, instead of the opposite.

When E is reflexive, then all the shorthands from the bra-ket notation can be used. That is, a dual element X can be denoted without ambiguity as $\langle \cdot, X \rangle_E$, and a primal element $x = x''$ as $\langle x, \cdot \rangle_E$. Moreover, for a bounded operator $T : E \to F$, we can write without ambiguity $\langle Tx, Y \rangle_F = \langle x | T | Y \rangle$. However since there are many interesting Banach spaces that are not reflexive, we will not use such shorthands.

1.2 Hahn-Banach theorem and useful consequences

Theorem 1.1 (Hahn-Banach theorem)**.** Let $\underline{E}$ be a linear subspace of a normed vector space $(E, \|\cdot\|)$ and let $f : \underline{E} \to \mathbb{R}$ be a bounded linear form on $(\underline{E}, \|\cdot\|)$.

Then there exists $g : E \to \mathbb{R}$ a bounded linear form on all of E, such that

- g is an extension of f: $g|_{\underline{E}} = f$;
- The extension "comes to no cost" in operator norm: $|||g||| = |||f|||$.

(Here the operator norms are with respect to their respective domains: $|||f||| = \sup_{x \in \underline{E}; \|x\| \leq 1} |f(x)|$, $|||g||| = \sup_{x \in E; \|x\| \leq 1} |g(x)|$.)

As one of the many important consequences of that theorem, we have the existence of norming functionals.

Definition 1.3 (Norming functional)**.** Let E be a Banach space.

For all $x \in E \setminus \{0_E\}$, there exists $X \in E'$ such that $X(x) = \langle x, X \rangle_E = \|x\|_E$ and $\|X\|_{E'} = 1$. X is then called a *norming functional* of x.

By convention, any $X \in B^{(E')}$ will be called a norming functional of 0_E.

Importantly, the norming functional X is not unique in general, and there is no generic way to construct it – the proof is not constructive.[1] (This stands in contrast with the case of Hilbert spaces, where the norming functional is unique and given by the Riesz representation theorem.)

Proposition 1.2. The primal E injects isometrically into the bidual E''.

Proof. To any $x \in E$, associate $x'' \in E''$ defined by $\forall X \in E', \langle X, x'' \rangle_{E'} = \langle x, X \rangle_E$.

To show that this linear mapping is injective, let $x \in E$ such that $x'' = 0_{E''}$, i.e $\langle x, X \rangle_E = 0$ for all $X \in E'$. Then $x = 0_E$; indeed otherwise we obtain a contradiction by taking X to be a norming functional of x.

To show that this mapping is isometric, let $x \in E$ and let us show that $\|x''\|_{E''} := \sup_{\|X\|_{E'} \leq 1} \langle x, X \rangle_E$ is equal to $\|x\|_E$.

[1] Or rather, to the pure functional analyst there is no satisfying way to construct a norming functional, but to the convex analyst there is... See below.

- For all $X \in E'$ such that $\|X\|_{E'} \leq 1$, it holds $\langle x, X\rangle_E \leq \|x\|_E$, so taking the sup over X yields $\|x''\|_{E''} \leq \|x\|_E$.
- Let X a norming functional of x. Then $\|x''\|_{E''} \geq \langle x, X\rangle_E = \|x\|_E$.

□

Another important consequence of the Hahn-Banach theorem is the following density criterion.

Lemma 1.3 (Hahn-Banach density criterion)**.** Let E be a Banach space. Let $\underline{E}$ be any subspace and A any subset.

$$\underline{E} \text{ is dense in } E \iff \underline{E}^{\perp} := \{X \in E';\ \forall x \in \underline{E}, \langle x, X\rangle_E = 0\} = \{0_{E'}\} \tag{1.1}$$

$$\text{span}(A) \text{ is dense in } E \iff A^{\perp} := \{X \in E';\ \forall a \in A, \langle a, X\rangle_E = 0\} = \{0_{E'}\} \tag{1.2}$$

The set $A^{\perp}$ is called the *annihilator* of A. Note that $A^{\perp} = (\text{span}(A))^{\perp}$. In the case where E is a Euclidean space, $A^{\perp}$ is just (up to isometry) the orthogonal complement of span(A).

Proof. The direct implication is trivial.

For the other direction, suppose to the contrary that $\underline{E}$ is not dense in E, and let $x_0 \in E \setminus \overline{\underline{E}}$. By definition there exists $\delta > 0$ such that $\|x - x_0\| \geq \delta$ for all $x \in \underline{E}$. Now the mapping

$$f : \begin{bmatrix} \underline{E} + \mathbb{R}x_0 \to \mathbb{R} \\ x + \lambda x_0 \mapsto \lambda \end{bmatrix} \tag{1.3}$$

is well-defined, and one can check that it is a bounded linear form with norm at most $1/\delta$.

Now let g an extension of f to all of E, by Hahn-Banach theorem. Then g contradicts the condition on the right-hand-side.

The second part of the lemma follows from the first by noting that $A^{\perp} = (\text{span}(A))^{\perp}$. □

2 Convex analysis (convex duality) cheat sheet

For the rest of this subsection, fix a Banach space E.

Definition 2.1. A function $f : E \to \mathbb{R} \cup \{+\infty\}$ is called *convex* if

$$\forall x, y \in E, \forall t \in [0,1],\ f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{2.1}$$

For any convex $f : E \to \mathbb{R} \cup \{+\infty\}$,

- The *domain* of f is the convex set $\text{dom}(f) = \{x \in E; f(x) < \infty\}$. f is called *proper* if $\text{dom}(f) \neq \varnothing$.
- f is called *lower-semicontinuous* (l.s.c) if its sub-level sets are closed, i.e for each $c \in \mathbb{R}$, $\{x \in E; f(x) > c\}$ is an open set.

Denote $\Gamma(E)$ the set of proper l.s.c convex functions over E.

Definition 2.2. For any proper convex function $f : E \to \mathbb{R} \cup \{+\infty\}$,

- The *subdifferential* of f at a point $x_0 \in E$ is the set

$$\partial f(x_0) = \left\{X \in E'; \forall x \in E, f(x) \geq f(x_0) + \langle x - x_0, X\rangle_E\right\}. \tag{2.2}$$

  f is called *subdifferentiable* at x_0 if $\partial f(x_0) \neq \varnothing$. Note that f can be subdifferentiable at x_0 only if $x_0 \in \mathrm{dom}(f)$.

- f is called *differentiable* at x_0 if $\partial f(x_0)$ is a singleton. Its unique element is then called the *differential* of f at x_0 and denoted $Df(x_0)$ or $\nabla f(x_0)$.

Note that the definitions of "subdifferential" and "differential" above look different from the usual ones, since they only apply to convex functions. It can be shown that our definitions are compatible with the usual ones from real analysis.[2]

Proposition 2.1. For any proper l.s.c function $f : E \to \mathbb{R} \cup \{+\infty\}$ such that $\mathrm{dom}(f)$ is open,

- f is convex iff $\mathrm{dom}(f)$ is convex and $\partial f(x_0) \neq \varnothing$ for all $x_0 \in \mathrm{dom}(f)$.
- If f is convex, then it is differentiable at x_0 (in the usual real-analytic sense) iff $\partial f(x_0)$ is a singleton, and the differential of f at x_0 (in the usual real-analytic sense) is then $Df(x_0)$.

2.1 Convex conjugate

Definition 2.3. For any proper function $f : E \to \mathbb{R} \cup \{+\infty\}$, the *convex conjugate* of f (a.k.a Fenchel-Legendre a.k.a Legendre-Fenchel a.k.a Fenchel a.k.a Legendre transform) is the function

$$f^* : \left[E' \to \mathbb{R} \cup \{+\infty\}, X \mapsto \sup_{x \in E} \langle x, X\rangle_E - f(x)\right]. \tag{2.3}$$

Proposition 2.2 (Fenchel-Moreau theorem)**.** For any proper function f, f^* is a proper l.s.c convex function.

A function f is a proper l.s.c convex function iff $f^{**} = f$.

For any proper function f, f^{**} is the tightest convex relaxation of f, in the sense that the epigraph of f^{**} is the convex hull of the epigraph of f. This is easy to visualize for functions over the real line.

Remark 2.1. In the proposition above, f^{**} is understood as a mapping from E to $\mathbb{R}$. Looking at the definitions, it would be more natural to view f^{**} as a mapping from E'' to $\mathbb{R}$ instead, which would be more general since E injects isometrically into E''. However in convex analysis we typically don't care about what happens outside of E.

More precisely: to be completely general and consistent with notation, we could define $f^{**} = (f^*)^*$ over E'' by

$$\forall z \in E'',\ f^{**}(z) = \sup_{X \in E'} \langle X, z\rangle_{E'} - f^*(X). \tag{2.4}$$

Since $\langle X, x''\rangle_{E'} = \langle x, X\rangle_E$ (where $[E \to E'', x \mapsto x'']$ denotes the canonical injection), the restriction of f^{**} to E is then – and this equation is typically taken as the definition of f^{**}:

$$\forall x \in E,\ f^{**}(x) = \sup_{X \in E'} \langle x, X\rangle_E - f^*(X). \tag{2.5}$$

[2]In other words, in this document we are only concerned with content covered in Rockafellar's 1970 "Convex Analysis" book, whereas in other contexts Rockafellar's 1998 "Variational Analysis" book may be a better reference.

Remark 2.2. In the context of convex analysis it is common to denote adjoint/conjugate/dual objects with a superscript "$*$". In other contexts that symbol connotes involution, which may be misleading. However for convex analysis there is not much risk of mistake, precisely because of the previous remark: we only ever care about what happens in E and E', never about the bidual space E''. In particular, even if f is not a proper l.s.c function, we may always write $(f^{**})^* = f^*$.

Accordingly, from here on we will follow the common practice and use x^* (instead of X) to denote a generic element of E'.

Many of the useful properties of convex conjugates can be found here: `https://en.wikipedia.org/w/index.php?title=Convex_conjugate&oldid=1007941296`.

2.2 Convex conjugates vs. subdifferentials

Proposition 2.3 (Fenchel-Young inequality)**.** Let $f \in \Gamma(E)$. By definition of the convex conjugate, we have Fenchel-Young's inequality:

$$\forall x \in E, \forall x^* \in E',\ f(x) + f^*(x^*) \geq \langle x, x^* \rangle_E \,. \tag{2.6}$$

For $x \in E$ and $x^* \in E'$,

- $x^* \in \partial f(x)$ iff x^* saturates Fenchel-Young's inequality, iff x^* achieves the sup in the definition of $f(x) = f^{**}(x) = \sup_{x^* \in E'} \langle x, x^* \rangle_E - f^*(x^*)$.
- $x \in \partial f^*(x^*)$ iff x saturates Fenchel-Young's inequality, iff x achieves the sup in the definition of $f^*(x^*) = \sup_{x \in E} \langle x, x^* \rangle_E - f(x)$.
- $x^* \in \partial f(x)$ iff $x \in \partial f^*(x^*)$.

Remark 2.3 (Subdifferentials as correspondence)*.* (Rockafellar 1970, Theorem 24.9) Up to an additive constant, $f \in \Gamma(E)$ is characterized by the binary relation $\mathcal{R}$ given by

$$x\mathcal{R}x^* \iff f(x) + f^*(x^*) = \langle x, x^* \rangle_E \,. \tag{2.7}$$

Proposition 2.4 (Norming functionals as subdifferentials)**.** The norm $\|\cdot\|_E$ is a proper continuous convex function by definition.

For any $x \in E$, $x^* \in E'$ is a norming functional for x iff $x^* \in \partial \|\cdot\|_E(x)$. In symbols,

$$x^* \in \partial \|\cdot\|_E(x) \iff \begin{cases} \|x^*\|_{E'} = 1 \\ \langle x, x^* \rangle = \|x\|_E \end{cases} \tag{2.8}$$

In particular, $\|\cdot\|_E$ is differentiable at x iff $\partial \|\cdot\|_E(x)$ is a singleton, iff x has a unique norming functional. If $\|\cdot\|_E$ is differentiable everywhere, then the mapping $[x \mapsto \nabla \|\cdot\|_E(x)]$ is well-defined and is called the *duality mapping*.

Thus, to the convex analyst, norming functionals are not a magical byproduct of the Hahn-Banach theorem, but simply a subgradient of the norm.[3]

[3]This explains why many papers on RKBSs make the assumption that the Banach space considered has a (Gateaux) differentiable norm: it ensures that the duality mapping is well-defined (single-valued), and gives an explicit way to compute it

2.3 Convex conjugacy swaps strict convexity for differentiability, and strong convexity for smoothness

Definition 2.4. Let $f \in \Gamma(E)$ and let $\mu > 0$, $L > 0$.

f is *strictly convex* if for all $x_0 \in \mathrm{dom}(f)$, there exists $g \in \partial f(x_0)$ such that the strict inequalities hold:

$$\forall x \in E \setminus \{x_0\},\ f(x) > f(x_0) + \langle x - x_0, g \rangle_E\,. \tag{2.9}$$

f is *μ-strongly convex* if for all $x_0 \in \mathrm{dom}(f)$, there exists $g \in \partial f(x_0)$ such that[4]

$$\forall x \in E,\ f(x) \geq f(x_0) + \langle x - x_0, g \rangle_E + \frac{\mu}{2} \|x - x_0\|_E^2\,. \tag{2.10}$$

f is differentiable everywhere if it is differentiable at each point of its domain, that is, if for all $x_0 \in \mathrm{dom}(f)$, there exists *a unique* $g \in E'$ such that

$$\forall x \in E,\ f(x) \leq f(x_0) + \langle x - x_0, g \rangle_E\,. \tag{2.11}$$

f is *L-smooth* if for all $x_0 \in \mathrm{dom}(f)$, there exists $g \in \partial f(x_0)$ such that

$$\forall x \in E,\ f(x) \leq f(x_0) + \langle x - x_0, g \rangle_E + \frac{L}{2} \|x - x_0\|_E^2\,. \tag{2.12}$$

Note that μ-strong convexity implies strict convexity, and that L-smoothness implies differentiability everywhere.

Proposition 2.5 (Kakade, Shalev-Shwartz, and Tewari, 2009)**.** Let $f \in \Gamma(E)$.

f is strictly convex iff f^* is differentiable.

f is μ-strongly convex iff f^* is $1/\mu$-smooth.

A Bonus functional analysis stuff

A.1 Canonical isometry between operators and bilinear forms

Definition A.1 (Dual operator)**.** For a bounded operator $T : E \to F$ between Banach spaces, its *dual (a.k.a adjoint) operator*[5] is $T' : F' \to E'$ defined by $\forall Y \in F', T'Y : x \mapsto \langle Tx, Y \rangle_F$. In other words,

$$\forall (x, Y) \in E \times F',\ \ \langle Tx, Y \rangle_F = \langle x, T'Y \rangle_E \tag{A.1}$$

Proposition A.1 (Specifying an operator by its action on the dual of the codomain)**.** To a bounded operator $T : E \to F$, associate the bilinear form

$$B_T : [(E \times F') \to \mathbb{R},\ (x, Y) \mapsto \langle Tx, Y \rangle_F = \langle x, T'Y \rangle_E]\,. \tag{A.2}$$

Let $\mathcal{B}_b(E \times F')$ denote the space of continuous bilinear forms over $E \times F'$, equipped with the norm

$$|||B|||_{E \times F'} = \sup_{\|x\|_E \leq 1} \sup_{\|Y\|_{F'} \leq 1} |B(x, Y)|\,. \tag{A.3}$$

[4]See also `https://xingyuzhou.org/blog/notes/strong-convexity`.

[5]Personally I prefer the term "dual operator", as it allows to stress that the bidual operator T'' is not the same as the primal in general; while the term "adjoint" connotes involution, which can be misleading here.

Then $\begin{bmatrix} (\mathcal{L}_b(E,F), |||\cdot|||_{E\to F}) \to (\mathcal{B}_b(E\times F'), |||\cdot|||_{E\times F'}) \\ T \mapsto B_T \end{bmatrix}$ is an injective isometric linear map between Banach spaces. It is bijective if $F'' = F$.

In words, any operator $T : E \to F$ is fully characterized by the action of Tx on F' for each $x \in E$, i.e by the bilinear form B_T. Furthermore, B_T is continuous with $|||B_T|||_{E\times F'} = |||T|||_{E\to F}$. However an arbitrary bilinear form $B \in \mathcal{B}_b(E\times F')$ may not induce an operator $T : E \to F$ in general.

Proof. Everything is trivial, except perhaps for the completeness of $(\mathcal{B}_b(E\times F'), |||\cdot|||_{E\times F'})$, for which we refer to `https://math.stackexchange.com/questions/185103/completeness-of-the-space-of-bounded-bilinear-maps`.

For injectiveness and isometry, use the existence of norming functionals (Hahn-Banach theorem).

For surjectiveness when F is reflexive, notice that any $B \in \mathcal{B}_b(E\times F')$ defines an operator $T : E \to F''$ by $\forall Y \in F',\ \langle Y, Tx\rangle_{F'} = B(x, Y)$. □

Proposition A.2 (Specifying an operator by its action on a predual of the codomain)**.** Suppose that there exists a Banach $\widetilde{F}$ such that $F = (\widetilde{F})'$, i.e $\widetilde{F}$ is *a predual space* of F.

To a bounded operator $T : E \to F$, associate the bilinear form

$$\widetilde{B}_T : \left[E \times \widetilde{F} \to \mathbb{R}, (x, z) \mapsto \langle z, Tx\rangle_{\widetilde{F}}\right]. \tag{A.4}$$

Let $\mathcal{B}_b(E\times\widetilde{F})$ denote the space of continuous bilinear forms over $E\times\widetilde{F}$, equipped with the norm $\left|\left|\left|\widetilde{B}\right|\right|\right|_{E\times\widetilde{F}}$ defined similarly as in the previous proposition.

Then $\begin{bmatrix} (\mathcal{L}_b(E,F), |||\cdot|||_{E\to F}) \to (\mathcal{B}_b(E\times \widetilde{F}), |||\cdot|||_{E\times \widetilde{F}}) \\ T \mapsto \widetilde{B}_T \end{bmatrix}$ is an isometric isomorphism between Banach spaces.

In words, compared to the previous proposition: any operator $T : E \to F$ is also fully characterized by the action of Tx on $\widetilde{F}$ for each $x \in E$, i.e by $\widetilde{B}_T$; and $\widetilde{F}$ is "small enough" so that conversely, any arbitrary bilinear form $\widetilde{B} \in \mathcal{B}_B(E\times\widetilde{F})$ induces an operator $T : E \to F$.

Remark A.1. Not all Banach spaces have a predual space. See e.g `https://math.stackexchange.com/questions/137677/what-is-the-predual-of-l1`.

B Bonus convex analysis stuff

B.1 Bregman divergence

Definition B.1. Let $\psi \in \Gamma(E)$ strictly convex and differentiable everywhere, i.e $\mathrm{dom}(\psi) = E$ and $\partial\psi$ and $\partial\psi^*$ are both single-valued. The Bregman divergence with respect to ψ is defined as

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla\psi(y)\rangle = \psi(x) - [\mathrm{Linearize}_y \psi](x) \tag{B.1}$$

Remark B.1. A possible interpretation: note that

$$\psi^*(x^*) = \sup_{t\in\mathbb{R}} t \quad \text{s.t} \quad \forall x \in E, \psi(x) \le \langle x, x^*\rangle - t \tag{B.2}$$

and that

$$D_{\psi}(x, y) = \psi(x) + \psi^*(y^*) - \langle x, y^* \rangle = \psi(x) + (\psi(\cdot - x))^* (y^*) \tag{B.3}$$

$$D_{\psi^*}(x^*, y^*) = \psi^*(x^*) + \psi(y) - \langle y, x^* \rangle = \psi^*(x) + (\psi^*(\cdot - x^*))^* (y) \tag{B.4}$$

where $y^* = \nabla\psi(y)$ i.e $y = \nabla\psi^*(y^*)$.

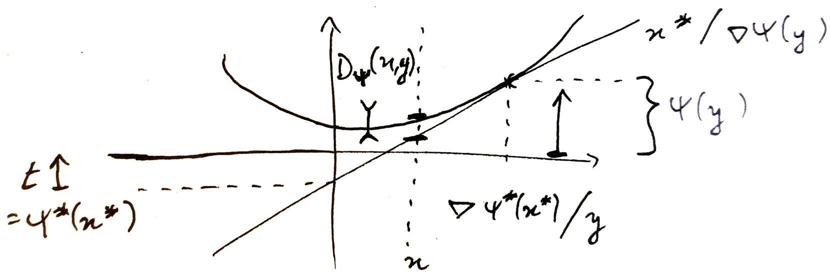

Figure 1: Bregman divergence

Proposition B.1. Let $\psi \in \Gamma(E)$ strictly convex and differentiable everywhere, and let any $w_0 \in E$. Denote $\varphi = D_{\psi}(\cdot, w_0)$. Note that $\varphi \in \Gamma(E)$ and is strictly convex and differentiable everywhere.

Then for any $w, w' \in E$, $D_{\varphi}(w, w') = D_{\psi}(w, w')$ independent of w_0.

B.2 Proximal operator

In this subsection, let E be a Hilbert space, which is a Banach space for which the dual is isometrically isomorphic to the primal.

Definition B.2. For any convex $f : E \to \mathbb{R}$, define the proximal operator $\text{prox}_f : E \to E$ as

$$\forall x \in E, \text{prox}_f(x) = \underset{u \in E}{\arg\min}\, f(u) + \frac{\|u - x\|^2}{2} \tag{B.5}$$

Note that the function appearing in the arg min is 1-strongly convex in u so that the arg min is indeed unique.

The proximal operator has nice properties, is computable in closed form for a number of interesting choices of f, and is used in many optimization algorithms; see `https://en.wikipedia.org/w/index.php?title=Proximal_gradient_methods_for_learning&oldid=1028229614`.