The Ubiquitous Subgradient

Outline: Subgradients are generalized gradients of convex functions. While you can't minimize $\varphi(x) = |x|$ by setting its derivative equal to zero (because the derivative doesn't exist at the minimizer), you *can* minimize it by setting its *subgradient* equal to zero. Subgradients free us from worrying about differentiability — at least for convex functions.

We have already seen applications of subgradients to minimization[1]. There are other applications of subgradients as well. These notes concentrate on the following three points:

1) Subgradients are ubiquitous, and so, therefore, are minimization problems.
2) The "usual" procedure (minimize a given a function by setting its derivative to zero) can be reversed: when a subgradient appears in a problem, use its (convex) antiderivative to aid analysis.
3) Subgradients are special **monotone** operators — operators that generalize the nonnegative real numbers or positive semi-definite matrices — in the same sense that *symmetric* positive semi-definite matrices are special matrices.

Definitions and Examples:

Definition: A function $\varphi(x) : X \to R$ is **convex** if for all $\lambda \in [0, 1]$ and all x and $y \in X$,

$$\varphi((1-\lambda)x + \lambda y) \leq (1-\lambda)\varphi(x) + \lambda\varphi(y). \tag{Def$_\varphi$}$$

It is important to "see" that this definition means that graphs of convex functions lie below secant lines connecting points on the graph. Please draw a picture to convince yourself.

Example: Squares of all kinds are convex: since $\lambda \in [0, 1]$,

$$\begin{aligned}\big[(1-\lambda)x + \lambda y\big]^2 &= (1-\lambda)x^2 + \lambda y^2 - \underbrace{\lambda(1-\lambda)(x-y)^2}_{\geq 0} \\ &\leq (1-\lambda)x^2 + \lambda y^2\end{aligned}$$

The expansion above could just as well have been in R^n or, indeed, any (semi-)inner product space:

$$\begin{aligned}\|(1-\lambda)x + \lambda y\|^2 &= (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - \lambda(1-\lambda)\,\|x-y\|^2 \\ &\leq (1-\lambda)\|x\|^2 + \lambda\|y\|^2.\end{aligned}$$

Note in particular that the $\|\cdot\|$ need only be *semi*-definite. For example, in R^n

$$\varphi(x) = x^T A x$$

is convex precisely when A is positive *semi*-definite:

$$\begin{aligned}\left((1-\lambda)x + \lambda y\right)^T A\left((1-\lambda)x + \lambda y\right) &= (1-\lambda)x^T Ax + \lambda y^T Ay - \underbrace{\lambda(1-\lambda)(x-y)^T A(x-y)}_{\geq 0} \\ &\leq (1-\lambda)x^T Ax + \lambda y^T Ay.\end{aligned}$$

Remark: The tight estimate

$$\|(1-\lambda)x + \lambda y\|^2 = (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - \lambda(1-\lambda)\,\|x-y\|^2$$

[1] Mau Nam Nguen, *The Weiszfeld Algorithm in the Light of Convex Analysis*

is a "generalized parallelogram law" for $\varphi(x) = \|x\|^2$. People say things like, "The balls in Hilbert space are especially round (or uniformly convex)" when referring to estimates of this type.

Example: (Unsquared) seminorms are convex, even if they don't come from inner products:

$$\begin{aligned} \|(1-\lambda)x + \lambda y\| &\leq \|(1-\lambda)x\| + \|\lambda y\| && \text{by the triangle inequality} \\ &= (1-\lambda)\|x\| + \lambda\|y\| && \text{by "positive scalar homogeniety".} \end{aligned}$$

In particular, $\varphi(x) = |x|$ on R and $\varphi(f) = \left(\int |f(x)|^p \, dx\right)^{1/p}$ on L^p are convex. Even

$$\varphi(f) = \sqrt{\int_\Omega |f'|^2 \, dx}$$

is convex.

Example: Powers $p \geq 1$ of convex functions are convex. Indeed, if g is any monotone (non-decreasing), convex function and f is convex, then the composition $g \circ f$ is convex:

$$\begin{aligned} f((1-\lambda)x + \lambda y) &\leq (1-\lambda)f(x) + \lambda f(y) && \text{since } f \text{ is convex, so} \\ g\Big(f((1-\lambda)x + \lambda y)\Big) &\leq g\Big((1-\lambda)f(x) + \lambda f(y)\Big) && \text{since } g \text{ is monotone} \\ &\leq (1-\lambda)g\left(f(x)\right) + \lambda g\left(f(y)\right) && \text{since } g \text{ is convex.} \end{aligned}$$

In particular, $\varphi(f) = \int |f|^p \, dx$ is convex for $p \geq 1$, as are the square semi-norms

$$\varphi_1(u(\cdot)) = \int_0^1 |u'(x)|^2 \, dx \qquad \text{and} \qquad \varphi_n(u(\cdot)) = \int_X |\nabla u(x)|^2 \, dx$$

on appropriate function spaces. If the function space contains any constant functions, then the minimum of $\varphi_n()$ is zero, and all the constant functions are minimizers. In this case, $\sqrt{\varphi_n()}$ is a semi-norm.

Example: All affine functions

$$\varphi(x) = ax + b$$

are convex, as are their negatives.

Example: Sums of convex functions are convex. In particular,

$$\varphi(u(\cdot)) = \int_0^1 |u'(x)|^2 - u(x)f(x) \, dx$$

is convex since it is the sum of a convex and a linear functional. What are the minimizers of φ?

Subgradients: Let $h = y - x$ in the definition of convexity (Def_φ):

$$\varphi(x + \lambda h) \leq \varphi(x) + \lambda\left[\varphi(x+h) - \varphi(x)\right] \quad \forall\, \lambda \in [0,1].$$

If $\lambda > 0$, then the inequality

$$\frac{\varphi(x+\lambda h) - \varphi(x)}{\lambda} \leq \varphi(x+h) - \varphi(x)$$

says that the difference quotients are increasing in the differences — larger increments h correspond to larger difference quotients than do smaller increments λh. Please draw a picture of a convex function to "see" that this interpretation is correct.

If the directional derivative of $\varphi()$ in the direction of h exists, then letting $\lambda \to 0^+$ gives the estimate

$$D_h\varphi(x)h \leq \varphi(x+h) - \varphi(x)$$

If we put $y = x + h$, then

$$\varphi(x) + D_h\varphi(x)(y-x) \leq \varphi(y)$$

The left side is just Taylor's linear approximation of $\varphi(x)$ about the point x, and its graph is the graph of a tangent line in the direction of $h = y - x$. In English, then, the inequality says *graphs of convex functions lie above their tangent lines.* Convex functions like $\varphi(x) = |x|$ may fail to have derivatives in the classical sense, but if we turn our inequality into a definition, they *will* have subgradients:

Definition: We say that $\partial\varphi(x)$ is a **subgradient** of (the convex function) $\varphi : D \to R$ at x if

$$\varphi(y) \geq \varphi(x) + \partial\varphi(x)\big(y - x\big) \quad \forall\, y \in D. \tag{Def$_{\partial\varphi}$}$$

If the domain of φ is (a subset of) R, then $\partial\varphi(x)$ is just a real number. If the domain of φ is (a subset of) R^n, then $\partial\varphi(x)$ is an n-dimensional row vector representing the gradient (with respect to whatever basis we've chosen). In general, however, if the domain of φ is a (convex subset of a) vector space X, then $\partial\varphi(x)$ belongs to the dual vector space X^*. In this general case, the "linear term" $\partial\varphi(x)\big(y-x\big)$ means "apply the functional $\partial\varphi(x) \in X^*$ to the increment $y - x \in X$.

Example: The subgradient of a (perfect) square is twice the identity:

$$\begin{aligned} y^2 &= x^2 + \ 2x\,(y-x) + (y-x)^2 \\ &\geq x^2 + \underbrace{2x}_{\partial x^2}(y-x), \end{aligned}$$

so $\partial\varphi(x) = 2x$ is a subgradient of $\varphi(x) = x^2$. The argument is identical in any semi-inner product space:

$$\begin{aligned} \|y\|^2 &= \|x\|^2 + 2\langle x, y-x\rangle + \|y-x\|^2 \\ &\geq \|x\|^2 + \langle \underbrace{2x}_{\partial\|x\|^2}, y-x\rangle, \end{aligned}$$

where $\langle u, v\rangle$ is the semi-inner product of u and v. This estimate says that, by definition, $\partial\varphi(x) = 2x$ is a subgradient of $\varphi(x) = x^2$.

Square semi-norms are used in interesting algorithms. Suppose A is an $n \times n$ symmetric, positive semi-definite matrix, and define the semi-inner product on R^n by

$$\langle u, v\rangle_A = u^T A v. \tag{$\langle\cdot,\cdot\rangle_A$}$$

Let us trace through the computation

$$\begin{aligned} \|y\|^2 &= x^T A x + 2x^T A y + (y-x)^T A (y-x) \\ &\geq x^T A x + 2x^T A y \\ &= \|x\|^2 + (\ \underbrace{2Ax}_{\partial\|x\|^2\,?}\ )^T y \end{aligned}$$

Is it contradictory that $\partial\|x\|^2 = 2A$, not $2I$? No. The ordinary Euclidean inner product $x^T y$ is not the same as the inner product $\langle x, y\rangle_A$. A change of basis can convert the latter to the former, but we don't have to

work that hard. By inspection, the map $2x \mapsto 2Ax$, which is sometimes called the **Riesz** mapping, takes the subgradient with respect to the (semi-)inner product $\langle \cdot, \cdot \rangle_A$ to the subgradient with respect to the Euclidean inner product. Algorithms using the inner product $\langle \cdot, \cdot \rangle_A$ are called **conjugate gradient** algorithms. "The" Conjugate Gradient Algorithm is the Gram-Schmidt orthogonalization process (= QR factorization) using the inner product $\langle \cdot, \cdot \rangle_A$.

Example: An important Riesz mapping arises from the squared semi-norm

$$\varphi(f) = \|f\|^2 = \int_0^1 \left(f'(x)\right)^2 \, dx.$$

The corresponding semi-inner product is

$$\langle f, g \rangle = \int_0^1 f'(x) g'(x) \, dx.$$

We know the subgradient of $\varphi(f) = \langle f, f \rangle$ is $2I$, but it's the Riesz mapping that is interesting. In particular, we would like to express the results using g, not g'. All arguments of this type boil down to integration by parts to remove the derivative from g':

$$\langle f, g \rangle = f'(x) g(x) \Big|_0^1 - \int_0^1 f''(x) g(x) \, dx$$

There is always some restriction placed on the boundary values of f and g — these are boundary-value problems, after all — that makes the boundary terms go to zero. The simplest restriction is the *homogeneous* boundary condition $g(0) = 0 = g(1)$. Then

$$\partial\varphi(f)(g) = \int_0^1 \underbrace{-f''(x)}_{\partial \|f\|^2} g(x) \, dx$$

Students of the Calculus of Variations talk about the **variation** $\delta\varphi(f) = -f''$ and the corresponding **Euler-Lagrange** equations $\delta\varphi(f) = 0$, but seldom consider the inner product space. The Riesz mapping makes the inner product spaces an explicit part of the problem.

The (very important) Riesz map

$$\partial \frac{1}{2} \| \cdot \|^2 = -\frac{d^2}{dx^2}$$

is the negative of the **Laplacian**. In n-dimensions, the appropriate integration by parts formula is Gauss's Theorem (a.k.a. the Divergence Theorem) and the Laplacian is the differential operator

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots \frac{\partial^2}{\partial x_n^2}.$$

Mathematicians write the operator as Δ, but other disciplines use the notations

$$\Delta = \operatorname{div} \nabla = \operatorname{div} \operatorname{grad} = \nabla \cdot \overrightarrow{\nabla}$$

Our results say that the Riesz mapping of the inner product space with inner product

$$\langle f, g \rangle = \int_\Omega \nabla f(x) \left(\nabla g(x)\right)^T \, dv$$

into L^2 is the (negative of the) Laplacian, and that this operator is a subgradient.

Square norms are important, but so are un-squared norms:

Example: Let

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0, \\ [-1,1] & x = 0, \\ -1 & x < 0 \end{cases}$$

be the multi-valued "function" called "signum". Then

$$|y| \geq |x| + \underbrace{\operatorname{sgn}(x)}_{\partial|x|}(y - x)$$

no matter which choice of $\operatorname{sgn}(x)$ we make. Consequently, "the" subgradient of $|x|$ can be any of the choices

$$\partial|x| = \operatorname{sgn}(x).$$

This important example illustrates how subgradients can be multivalued. The collection of all subgradients at a point is the subdifferential:

Definition: The **subdifferential** of the convex function φ at x is the collection of all subgradients $\varphi(x)$.

Can we extend the real line's absolute value to a vector space's norm? Yes:

Example: The Hahn-Banach Theorem is written so that semi-norms have "duality mappings": Every $x \in X$ has an $x^* \in X^*$ with $\|x^*\|_{X^*} \leq 1$ and

$$x^*(x) = \|x\|$$

The definition of "$\|x^*\|_{X^*} \leq 1$" means that

$$\|y\| \geq x^*(y),$$

so the condition $x^*(x) = \|x\|$ means

$$\|y\| \geq \|x\| + \underbrace{x^*}_{\partial\|x\|}(y - x).$$

In English: every duality mapping is a subgradient of the semi-norm.

N.B.: If $f = 0$, any element x^* of norm ≤ 1 will serve as a duality mapping:

$$\|y\| \geq x^*(y) = \|0\| + x^*(y - 0).$$

What is the subgradient doing in the Hahn-Banach Theorem? In other words, what convex function does the subgradient minimize? For a fixed $x \in X$, define $\varphi_x : X^* \to R$ by

$$\varphi_x(y) = \begin{cases} \|x\| - y(x) & \text{if } \|y\|_{X^*} \leq 1 \\ +\infty & \text{if } \|y\|_{X^*} > 1 \end{cases}$$

Then $\varphi_x(\cdot)$ is convex because it is linear on the (convex) unit ball of X^*, and the infinite values outside the ball guarantee

$$\varphi_x\big((1-\lambda)y_1 + \lambda y_2\big) \leq (1-\lambda)\varphi_x(y_1) + \lambda\varphi_x(y_2)$$

even when one of the y_i is outside the unit ball. Furthermore, $\varphi_x(y) \geq 0$. In general, convex functions $\varphi(y) \geq 0$ that go to infinity with $\|y\|$ "ought" to have minima, at least in "complete" spaces. The Hahn-Banach theorem guarantees that a minimizer $\hat{y}$ does, indeed, exist:

$$\begin{aligned} & & \varphi_x(\hat{y}) \leq \varphi_x(y) & \quad \forall\, y \in X^* & \iff \\ \hat{y} \in B_{X^*} & \text{ and } & \|x\| - \hat{y}(x) \leq \|x\| - y(x) & \quad \forall\, y \in B_{X^*} & \iff \\ \|\hat{y}\|_{X^*} \leq 1 & \text{ and } & \hat{y}(x) \geq y(x) & \quad \forall\, y \in B_{X^*} & \iff \\ \|\hat{y}\|_{X^*} \leq 1 & \text{ and } & \hat{y}(x) \geq \|x\| & \quad \text{and } \hat{y} \text{ exists by the Hahn-Banach Theorem.} & \end{aligned}$$

Can we use the duality mapping to gain insight into the dual space? Yes. For example, in L^p, the separation

$$\|f\|_{L^p} = \frac{\int |f(x)|^p \, dx}{\|f\|_{L^p}^{p-1}} = \int \frac{\operatorname{sgn}(f(x))|f(x)|^{p-1}}{\|f\|_{L^p}^{p-1}} \cdot f(x) \, dx$$

suggests that the duality functional of f is

$$f^*(x) = \frac{\operatorname{sgn}(f(x))|f(x)|^{p-1}}{\|f\|_{L^p}^{p-1}}.$$

Where does $f^*(x)$ "live"? If the dual space to L^p is a normed space, then the norm of f^* should be 1 (as long as f is not zero a.e.). What norm must the dual space have if this functional has norm 1? We are only guaranteed that $\|f\|_{L^p}$ is finite, so the norm of f^* *must* be something in terms of the L^p norm of f. To get an L^p norm, integrate the q power of f^* (and ignore the denominator):

$$\left(\int |f(x)|^{(p-1)q} \, dx \right)^{\frac{1}{q}} \quad \text{must be a function of } \|f\|_{L^p}.$$

Consequently, q must satisfy $(p-1)q = p$ — in agreement with the identification of $L^{p*} = L^q$. In English: the duality functional helped us identify the dual space of L^p.

Minimization: Subgradients are the "right" tools to use when minimizing convex functions. From its definition ($\text{Def}_{\partial\varphi}$), we see that

$$\varphi(y) \geq \varphi(x) \quad \forall\, y \iff 0 \in \partial\varphi(x).$$

We use (the results of) minimization all the time, often without thinking about it. Even the simplest example contains the germs of many fruitful ideas. The next few examples illustrate "real world" applications of minimization that are easily understood (even without subgradients).

Example: Suppose you give a student 3 quizzes, and want to summarize the scores using a single number. What do you do? Why?

Answer: Teachers often average the quiz scores q_1, q_2, and q_3 to get the single number

$$\bar{q} = \frac{q_1 + q_2 + q_3}{3}.$$

The average minimizes the sum-of-squares

$$E(q) = (q - q_1)^2 + (q - q_2)^2 + (q - q_3)^2,$$

so one reason we use the average is that it solves of the simultaneous equations

$$\begin{aligned} q &= q_1 \\ q &= q_2 \\ q &= q_3 \end{aligned}$$

"in the least-squares sense". We *could* teach high school students this fact when we teach them to complete the square:

$$\begin{aligned} E(q) &= 3q^2 - 2q(q_1 + q_2 + q_3) + \left[q_1^2 + q_2^2 + q_3^2\right] \\ &= 3\left[\left(q - \frac{q_1 + q_2 + q_3}{3}\right)^2 + \frac{q_1^2 + q_2^2 + q_3^2}{3} - \left(\frac{q_1 + q_2 + q_3}{3}\right)^2\right] \end{aligned}$$

Since the leftmost term inside the square brackets is non-negative, the minimum occurs when $q = \frac{q_1+q_2+q_3}{3}$.

Example: (continued:) What if one quiz is 50% of the grade and the other two are 25% each?

Answer: Solutions to *weighted* least squares problems are *weighted* averages. For example, if the first two quizzes each contribute 25% to the quiz grade, and the third quiz contributes 50%, the least squares function becomes

$$\begin{aligned} E_w(q) &= 0.25(q-q_1)^2 + 0.25(q-q_2)^2 + 0.5(q-q_3)^2 \\ &= q^2 - 2q(0.25q_1 + 0.25q_2 + 0.5q_3) + \left[0.25q_1^2 + 0.25q_2^2 + 0.5q_3^2\right] \\ &= (q - (0.25q_1 + 0.25q_2 + 0.5q_3))^2 + \left[0.25q_1^2 + 0.25q_2^2 + 0.5q_3^2\right] - (0.25q_1 + 0.25q_2 + 0.5q_3)^2 . \end{aligned}$$

The minimizer is the weighted average

$$\bar{q}_w = 0.25q_1 + 0.25q_2 + 0.5q_3.$$

Asides: Probabilists and statisticians would tell us to divide $E(q)$ by 3 and use the *mean* sum of squares. Then the completed square becomes

$$\frac{(q-q_1)^2 + (q-q_2)^2 + (q-q_3)^2}{3} = \left(q - \frac{q_1+q_2+q_3}{3}\right)^2 + \frac{q_1^2+q_2^2+q_3^2}{3} - \left(\frac{q_1+q_2+q_3}{3}\right)^2$$

When $q = \bar{q}$, the first term on the right drops out, and we have the Pythagorean identity

$$\underbrace{\frac{(q_1-\bar{q})^2 + (q_2-\bar{q})^2 + (q_3-\bar{q})^2}{3}}_{S^2} = \underbrace{\frac{q_1^2+q_2^2+q_3^2}{3}}_{\Sigma^2} - \underbrace{\left(\frac{q_1+q_2+q_3}{3}\right)^2}_{\bar{q}^2}$$

or

$$S^2 + \bar{q}^2 = \Sigma^2$$

The S^2 is (nearly) the statisticians' *sample variance*, and is often denoted by V. The Pythagorean identity suggests that some sort of *orthogonal projection* must be taking place.

Engineers will take the square root of the statisticians' S^2 and call the resulting S the *root mean square* (RMS) error.

The point: The more names an object such as $E(q)$ has, the more important it is likely to be. Every "real world" discipline has its own version of least squares problems, with its own nomenclature.

Medians: What is wrong with using the arithmetic mean to represent a data set? Outliers. The classic example is the average wealth of 10 people, one millionaire and 9 paupers. Their average wealth is

$$\frac{\$1,000,000 + 0 + 0 + \cdots + 0}{10} = \$100,000 \text{ per person,}$$

but this summary of the group's wealth is dissatisfyingly optimistic. We know how this happens: in the sum of squares

$$E(q) = (10^6 - q)^2 + (0-q)^2 + \cdots + (0-q)^2,$$

every deviation is squared, so the "outlier's" — the millionaire's — contribution skews the mean significantly. Economists (and freshmen) know what to do: minimize the sum of absolute values instead:

$$\varphi(q) = |10^6 - q| + |0-q| + \cdots + |0-q|$$

Question: How do we teach our high school students to "complete the absolute value" to minimize $\varphi(\cdot)$?

Answer: We don't. In fact, a huge (hidden) factor in the popularity of least squares is the fact that

The minimum of a *quadratic* function satisfies a *linear* equation.

That's why we call the class of problems "linear least squares", and they are popular in part because all the power of linear algebra is at our disposal to solve them.

So what do we do? We can still teach high school students to minimize the sum of absolute values in the one-dimensional problem. To minimize

$$\varphi(x) = |x - x_1| + |x - x_2| + \cdots + |x - x_n|,$$

order the data set so that $x_1 \le x_2 \le \cdots \le x_n$. To find the minimium, start with $x = x_1$ and move to the right. For $x \in [x_1, x_2]$, the sum of the first two terms is constant:

$$|x - x_1| + |x - x_2| = x_2 - x_1 \qquad \text{for } x \in [x_1, x_2]$$

Consequently, $\varphi(x)$ decreases as x increases because the remaining terms, $|x - x_k|$ for $x > 2$, decrease as x increases. Likewise, for $x \in [x_2, x_3]$, the sum of the first 4 terms is constant:

$$|x - x_1| + |x - x_2| + |x - x_3| + |x - x_4| = x_4 + x_3 - x_2 - x_1 \qquad \text{for } x \in [x_2, x_3]$$

Again, φ decreases as x increases because the remaining terms, $|x - x_k|$ for $x > 4$, decrease as x increases. This analysis continues until x reaches the middle of the data set. If the data set has an odd number of values, then the middle of the data set is the unique minimizer, and is called the **median**. If the data set has an even number of points, then the "middle" may not be a single point, but all the minimizers are called medians. For example, any point between 2 and 5 is a median of $\{1, 2, 5, 17\}$. We have solved the one-dimensional "least absolute values" problem. (In particular, a group consisting of one millionaire and 9 paupers has median wealth of \$0.)

Higher Dimensional Problems: Least-squares problems in many dimensions are not much harder than least-squares in 1 dimension. Common notations for square norms are $\|x\|^2 = x^T x$ in R^n and, more generally, $\|x\|^2 = \langle x, x\rangle$ in an inner product space. With the inner-product notation,

$$\begin{aligned}
\varphi(x) &= \|x - a_1\|^2 + \|x - a_2\|^2 + \cdots + \|x - a_n\|^2 \\
&= n\|x\|^2 - 2\langle a_1 + a_2 + \cdots + a_n,\, x\rangle + \|a_1\|^2 + \|a_2\|^2 + \cdots + \|a_n\|^2 \\
&= n\left\|x - \frac{a_1 + a_2 + \cdots + a_n}{n}\right\|^2 + n\left[\frac{\|a_1\|^2 + \|a_2\|^2 + \cdots + \|a_n\|^2}{n} - \left\|\frac{a_1 + a_2 + \cdots + a_n}{n}\right\|^2\right]
\end{aligned}$$

The function is minimized when $x = \bar{a} = \frac{a_1 + a_2 + \cdots + a_n}{n}$.

If each term in the average is weighted by λ_i (instead of by 1), then the corresponding function to minimize is

$$\begin{aligned}
\varphi(x) &= \lambda_1\|x - a_1\|^2 + \lambda_1\|x - a_2\|^2 + \cdots + \lambda_1\|x - a_n\|^2 \\
&= \left(\sum \lambda_i\right)\|x\|^2 - 2\langle \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n,\, x\rangle + \lambda_1\|a_1\|^2 + \lambda_2\|a_2\|^2 + \cdots + \lambda_n\|a_n\|^2 \\
&= \sum \lambda_i \left\|x - \frac{\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n}{\sum \lambda_i}\right\|^2 \\
&\quad + \sum \lambda_i \left[\frac{\lambda_1\|a_1\|^2 + \lambda_2\|a_2\|^2 + \cdots + \lambda_n\|a_n\|^2}{\sum \lambda_i} - \left\|\frac{\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n}{\sum \lambda_i}\right\|^2\right]
\end{aligned}$$

In English: Every weighted average minimizes a quadratic "error" function.

Remark: This is what Kuhn[2] means when he says:

> This says that $T(P)$ is the center of gravity of weights $\frac{w_i}{d_i(P)}$ placed at the vertices A_i. Hence, by elementary calculus, $T(P)$ is the unique minimum of the strictly convex function
>
> $$g(Q) = \sum_i \frac{w_i}{d_i(P)} d_i^2(Q).$$

The physicist turns Kuhn's "center of gravity of weights" into an experiment by balancing a (weightless) plate with weights λ_i at position a_i. The point of balance minimizes the

$$\text{Moment of Inertia relative to } x = \sum \lambda_i \|a_i - x\|^2$$

Question: What if the physicist (Torricelli, say) were to tie a bunch of strings together at a common knot, then tie masses λ_i to the ends of the strings, and drop the weights through holes at (vector) positions a_i drilled through a table? Where would the common knot come to rest?

Hint: Torricelli argued that the knot assumes a position where the potential energy of the system is a minimum. The potential energy is the sum of the height from the floor of each weight times its weight:

$$E = \sum_j h_j w_j.$$

Since the total length l_j of each string is constant,

$$E = \sum_j l_j w_j - \sum_j (l_j - h_j) w_j.$$

The first term on the right is constant, independent of the knot's position. The second is the sum of the weighted lengths $l_j - h_j$ — the lengths of the strings *on top of* the table. Minimizing the potential energy is therefore the same problem as minimizing the sum of the weighted distances from the holes to the knot.

Answer: The knot comes to rest where the function

$$\varphi(x) = \sum w_i \|x - a_i\|$$

achieves its minimum. Note that the norms are not squared — if they were, the problem would be an easy-to-solve least-squares problem, and Fermat would not have his name associated with it. Instead, the potential energy of the system is minimized when the weights are hanging as far below the table as possible, which means the string left on the table (the weighted sum above) should be as short as possible. This is the Fermat-Torricelli problem, and is much harder to solve than the least squares problem. We have already seen[3] how to use subgradients to solve the Fermat-Torricelli problem.

Σ-Δ Modulators: Subgradients appear in the "real world" — we just have to train our eyes to see them. When a problem has subgradient, it is natural to search for a related convex function.

Σ-Δ modulators are versatile electronic circuits. For example, Σ-Δ modulators are used in analog signal - to - digital signal conversion ("ADCs"). The analog signal is integrated (on a capacitor) over successive time intervals, giving a real-valued sequence f_n. The modulator accepts the real-valued f_n and, at every

[2] Kuhn, Harold, W., *A Note on Fermat's Problem*, Mathematical Programming, 4 (1973), p.102

[3] Mau Nam Nguen, *The Weiszfeld Algorithm in the Light of Convex Analysis*

iteration, computes a binary digit $\delta_n = \pm 1$ in a way that "represents" f_n. Binary digits are called **bits**, and a "digital" signal is one constructed from a finite number of bits.

The restriction that $\delta_n \in \{1, -1\}$ means that the modulator will not represent a function f_n well if $|f_n|$ is large. We therefore assume

$$\|f\|_{\ell^\infty} \leq 1.$$

The restriction that $\delta_n \in \{1, -1\}$ also means that we can only hope to approximate f_n using a combination of the δ_n — a running average, for example.

Example: Suppose $f_n = \frac{1}{2}$ for all n. The periodic sequence of period 4 defined by

$$\begin{aligned} \delta_0 &= 1 \\ \delta_1 &= -1 \\ \delta_2 &= 1 \\ \delta_3 &= 1 \\ &\vdots \end{aligned}$$

has the property that running averages over any 4 consecutive values exactly represent f_n.

We would like a simple electronic circuit that chooses the sequence δ_n from the values of f_n and the previous choices of δ_n. Let

$$x_{n+1} = x_n + f_n - \delta_n$$

be the "accumulated error" of the $f_n - \delta_n$. We would like to keep x_n small, at least on average. An obvious choice is, therefore,

$$\delta_n = \operatorname{sgn}(x_n) = \begin{cases} 1 & \text{if } x_n \geq 0 \\ -1 & \text{if } x_n < 0 \end{cases}$$

so that

$$x_{n+1} = x_n + f_n - \operatorname{sgn}(x_n) \tag{Σ-Δ}$$

(Note that the choice sgn(0) is arbitrary as far as the arithmetic is concerned, but must be ± 1 so the digital output is meaningful to the engineer.)

Example: For the constant $f_n = \frac{1}{2}$, the model (Σ-Δ) produces

$$\begin{array}{rcrcrcl} x_0 = & 0 & \text{so} & \delta_0 = & 1 & \text{and} & x_1 = 0 + \frac{1}{2} - 1 = -\frac{1}{2} \\ x_1 = & -\frac{1}{2} & \text{so} & \delta_1 = & -1 & \text{and} & x_2 = -\frac{1}{2} + \frac{1}{2} + 1 = 1 \\ x_2 = & 1 & \text{so} & \delta_2 = & 1 & \text{and} & x_3 = 1 + \frac{1}{2} - 1 = \frac{1}{2} \\ x_3 = & \frac{1}{2} & \text{so} & \delta_3 = & 1 & \text{and} & x_4 = \frac{1}{2} + \frac{1}{2} - 1 = 0, \end{array}$$

generating the periodic sequence of δ_i above.

Note that if we were to choose $\operatorname{sgn}(0) = -1$ instead of $\operatorname{sgn}(0) = +1$, then

$$\begin{array}{rcrcrcl} x_0 = & 0 & \text{so} & \delta_0 = & -1 & \text{and} & x_1 = 0 + \frac{1}{2} + 1 = \frac{3}{2} \\ x_1 = & \frac{3}{2} & \text{so} & \delta_1 = & 1 & \text{and} & x_2 = \frac{3}{2} + \frac{1}{2} - 1 = 1 \\ x_2 = & 1 & \text{so} & \delta_2 = & 1 & \text{and} & x_3 = 1 + \frac{1}{2} - 1 = \frac{1}{2} \\ x_3 = & \frac{1}{2} & \text{so} & \delta_3 = & 1 & \text{and} & x_4 = \frac{1}{2} + \frac{1}{2} - 1 = 0, \end{array}$$

producing the same periodic output (translated by 1) but producing a maximum x_k of $\frac{3}{2}$ — 50% larger than the maximum of the previous example. Similarly, the average value of x_n is $\frac{1}{4}$ in the first example, and $\frac{3}{4}$ in the second. The size of the intermediate x_n dictates constraints on the design of the modulator. Specifically, we have to build the modulator so that the x_n do not overflow. We would therefore like to prove stability results.

To gain some insights, consider the *differential* equation

$$\frac{d}{dt}x(t) = \varphi(t) - \operatorname{sgn}(x(t)). \tag{DE}$$

The $\operatorname{sgn}(x(t))$ is a subgradient — $\partial|x| = \operatorname{sgn}(x)$ — and we might seek other, similar, subgradients that are useful in our analysis. In particular, we might consider the family of subgradients

$$\operatorname{sgn}^+(x-c) = \begin{cases} 1 & x > c \\ 0 & x \le c. \end{cases}$$

The superscript $^+$ means "the positive part of", and the positive part of signum "function" is the subgradient of the ramp function:

$$\partial(x-c)^+ = \operatorname{sgn}^+(x-c)$$

because, for all $y \in R$,

$$\begin{aligned}(y-c)^+ &\ge (x-c)^+ + \operatorname{sgn}^+(x-c)\big(y-c-(x-c)\big)\\ &\ge (x-c)^+ + \operatorname{sgn}^+(x-c)\big(y-x\big).\end{aligned}$$

Multiply both sides of (DE) by $\operatorname{sgn}^+(x-c)$ and recognize the chain rule:

$$\operatorname{sgn}^+(x-c)\frac{d}{dt}x(t) = \frac{d}{dt}\left(x(t)-c\right)^+ = \operatorname{sgn}^+(x-c)\left(f(t) - \operatorname{sgn}(x(t))\right).$$

If $c > 0$, then either

1) $\operatorname{sgn}^+(x-c) = 0$ (and so, therefore, is the right side), or
2) $\operatorname{sgn}^+(x-c) \neq 0$, so $x > c > 0$, so $\operatorname{sgn}(x_n) = 1$.

Consequently,

$$\begin{aligned}\frac{d}{dt}\left(x(t)-c\right)^+ &= \operatorname{sgn}^+(x-c)\left(f(t)-1\right)\\ &\le 0 \quad \text{because } f(t) \le 1.\end{aligned}$$

The inequality means the convex function $(x(t)-c)^+$ is non-increasing. Such functions are called **Lyapunov** functions for the differential equation. The Lyapunov function provides the stability result we seek: if $x(t)$ starts out below c, it stays below c forever. The argument is one-sided, providing only an upper bound, but it works equally well to prove that the convex function $(x+c)^-$ is non-increasing, providing a lower bound.

Let us attempt the same argument on the difference equation (Σ-Δ). From the definition of the subgradient

$$\varphi(y) - \varphi(x) \ge \partial\varphi(x)(y-x)$$

and the computation of the subgradient of the ramp function,

$$\partial(x-c)^+ = \operatorname{sgn}^+(x-c)$$

we have, if $c > 0$

$$\begin{aligned}(x_{n+1}-c)^+ - (x_n-c)^+ &\ge \operatorname{sgn}^+(x_n-c)\left((x_{n+1}-c)-(x_n-c)\right)\\ &= \operatorname{sgn}^+(x_n-c)\left(f_n - \operatorname{sgn}(x_n)\right)\\ &= \operatorname{sgn}^+(x_n-c)\left(f_n-1\right)\\ &\le 0 \quad \text{because } f(t) \le 1.\end{aligned}$$

Too bad the subgradient inequality on the top line goes the wrong way! What shall we do?

Answer 1: Choose c large enough to guarantee that

$$\begin{aligned}(x_{n+1}-c)^+ - (x_n-c)^+ &\le \operatorname{sgn}^+(x_{n+1}-c)\left((x_{n+1}-c)-(x_n-c)\right)\\ &= \operatorname{sgn}^+(x_{n+1}-c)\left(f_n-\operatorname{sgn}(x_n)\right))\\ &= \operatorname{sgn}^+(x_{n+1}-c)\left(f_n-1\right)) \quad \text{if } c \text{ is large enough, so}\\ &\le 0.\end{aligned}$$

How large must we choose c? If $x_n > 0$, then

$$x_{n+1} = x_n + f_n - 1 > -\|f\|_{\ell^\infty} - 1$$

so

$$c = 1 + \|f\|_{\ell^\infty}$$

suffices. (Check: if $x_{n+1} > 1 + \|f\|_{\ell^\infty}$, then x_n could not have been negative, else x_{n+1} would have been $x_n + f_n + 1 < 1 + \|f\|_{\ell^\infty}$, not the other way around.)

N.B.: We have already seen that the upper bound can be (essentially) achieved.

Answer 2: Choose δ_n differently. If we could look into the future to set $\delta_n = \operatorname{sgn}(x_{n+1})$, then we would have the inequality we want:

$$\begin{aligned}(x_{n+1}-c)^+ - (x_n-c)^+ &\le \operatorname{sgn}^+(x_{n+1}-c)\left((x_{n+1}-c)-(x_n-c)\right)\\ &= \operatorname{sgn}^+(x_{n+1}-c)\left(f_n-\operatorname{sgn}(x_{n+1})\right))\\ &= \operatorname{sgn}^+(x_{n+1}-c)\left(f_n-1\right)) \le 0.\end{aligned}$$

How do we gain the clairvoyance? Mathematically, we can solve

$$x_{n+1} + \operatorname{sgn}\left(x_{n+1}\right) = x_n + f_n$$

for x_{n+1}:

$$x_{n+1} = \left(I + \operatorname{sgn}(\cdot)\right)^{-1}\left(x_n + f_n\right).$$

The inverse operator $\left(I + \operatorname{sgn}(\cdot)\right)^{-1}$ is the piecewise-linear function

$$\left(I + \operatorname{sgn}(\cdot)\right)^{-1} y = \begin{cases} y-1 & y > 1\\ 0 & y \in [-1,1]\\ y+1 & y < -1.\end{cases}$$

This means, for example, that if $x_n + f_n \in [-1,1]$, then

$$x_{n+1} = \left(I + \operatorname{sgn}(\cdot)\right)^{-1}\left(x_n + f_n\right) = 0$$

and the "bit" is

$$\delta_n = x_n + f_n \in \operatorname{sgn}(0) = [-1,1].$$

In other words, this elegant mathematical solution is an engineering disaster because it *violates the requirement that the output be* ± 1! Remember that the engineer requires $\delta_n \in \{-1,1\}$.

If we give up now, we will have earned the engineers' scorn. We therefore seek a compromise[4]. A smart choice is to choose $\delta_n = \pm 1$ so as to minimize the size of $x_{n+1} = x_n + f_n - \delta_n$. Then the mathematician and the engineer are satisfied.

[4] Stonick, J.T., Rulla, J.L., Ardalan, S.H., "Look-ahead decision-feedback Σ-Δ modulation", in *Acoustics, Speech, and Signal Processing* (ICASSP), vol.3 (1994), pp. 541–544.

Which choice of δ_n minimizes $|x_{n+1}|$?

$$\delta_n = \operatorname{sgn}(x_n + f_n) \tag{LADF}$$

With this choice of δ, the Σ-Δ modulator is (unfortunately) called a **Look Ahead, Decision Feedback** modulator. The stability estimate for the (LADF) modulator becomes

$$(x_{n+1} - c)^+ \le (x_n - c)^+ + (f_n - \operatorname{sgn}(x_n + f_n))\operatorname{sgn}^+(x_{n+1} - c),$$

and I claim c may be chosen to be 1. In fact, since $x - \operatorname{sgn}(x) > 1$, implies $x > 2$,

$$\text{if} \quad x_{n+1} = x_n + f_n - \operatorname{sgn}(x_n + f_n) > 1 \quad \text{then} \quad x_n + f_n > 2 \quad \text{and} \quad \delta_n = \operatorname{sgn}(x_n + f_n) = 1.$$

Consequently,

$$\begin{aligned}(x_{n+1} - 1)^+ &\le (x_n - 1)^+ + \operatorname{sgn}^+(x_{n+1} - 1)\big(f_n - \operatorname{sgn}(x_n + f_n)\big) \\ &= (x_n - 1)^+ + \operatorname{sgn}^+(x_{n+1} - 1)\big(f_n - 1\big) \\ &\le (x_n - 1)^+.\end{aligned}$$

This elegant stability result is better than the result for the standard Σ-Δ modulator, meaning the bounds on x_n are ± 1 instead of $\pm(1 + \|f\|)$.

The Point: The subgradient in the problem provided motivation for an improved architecture for the circuit. Convex functions motivated by the subgradient provide the stability analysis. Both the engineer and the mathematician are happy.

Monotonicity: We teach our calculus students that functions are convex ("concave up") when the first derivative is **monotone** (or non-decreasing). If the second derivative exists, it is therefore non-negative (positive semi-definite). The definition of subgradient means that

$$\begin{aligned}\varphi(y) - \varphi(x) &\ge \partial\varphi(x)(y - x), \text{ and} \\ \varphi(x) - \varphi(y) &\ge \partial\varphi(y)(x - y), \text{ so} \\ 0 &\ge \big(\partial\varphi(x) - \partial\varphi(y)\big)(y - x).\end{aligned}$$

Symmetrize the order of subtraction, and the result is

$$\big(\partial\varphi(y) - \partial\varphi(x)\big)(y - x) \ge 0.$$

We say that an operator $A : X \to X^*$ is **monotone** if

$$\big(A(y) - A(x)\big)(y - x) \ge 0 \tag{M}$$

Consequently, subgradients are monotone operators.

When $X = R$, definition (M) just says that A *preserves order*: if $x < y$, then $A(x) \le A(y)$. Definition (M) extends the notion of order to vector spaces of more than one dimension.

Every *linear* monotone operator in one dimension is a subgradient. In fact, every linear operator on R is just multiplication by a number a, and if the operator is monotone, then that number $a \ge 0$. The operator is therefore

$$A(x) = a \cdot x = \partial \underbrace{\left(\frac{a}{2}x^2\right)}_{\varphi(x)}.$$

In two or more dimensions, a *linear* monotone operator is multiplication by a semi-definite matrix. But there are two fundamental "flavors" of semi-definite matrices:

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = ax^2 + by^2 \geq 0 \text{ provided } a, b \geq 0, \text{ and}$$

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \text{ for any } a.$$

The symmetric matrix in the top example is a subgradient — indeed, it is the (Riesz mapping of the) subgradient of the square semi-norm $\|(x,\, y)\|^2 = \frac{a}{2}x^2 + \frac{b}{2}y^2$. The skew-symmetric matrix can not be a subgradient: if it were, its second derivative would have to be symmetric, and it isn't. We shall see later that the symmetric matrix is associated with the heat equation, and the skew-symmetric symmetric matrix is associated with the wave equation. We can use monotonicity methods to solve both the heat and wave equations, but the convergence rate for the the heat equation is faster than the rate for the wave equation. The subgradient explains the different rates of convergence.

Rates of Convergence: Our last topic is the initial value problem

$$\begin{cases} \dfrac{d}{dt}u(t) + A(u(t)) = 0 \\ \qquad\qquad\qquad u(0) = u_0 \end{cases} \tag{IVP}$$

The operator A may not be bounded (for example, $A = -\Delta$), but it is supposed to be monotone (see definition (M)). Examples of monotone linear operators A are

$A(u) = a \cdot u$ for some real $a \geq 0$ (so $A : R^1 \to R^1$),
$A(u) = \mathrm{A}u$ for some positive semidefinite matrix A (so $A : R^n \to R^n$),
$A(u) = -\Delta u$, so (IVP) is the heat equation, and

$$A\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) = \begin{bmatrix} 0 & -i\frac{\partial}{\partial x} \\ i\frac{\partial}{\partial x} & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \quad \text{so } (IVP) \text{ is the wave equation } u_{tt} + u_{xx} = 0 \text{ (and similarly for } v).$$

Recall that all subgradients are monotone.

To solve the initial value problem, replace the derivative in (IVP) with a difference quotient, and replace the *differential* equation with a *difference* equation:

$$\frac{u(t+h) - u(t)}{h} + A(u(t)) = 0, \quad \text{so}$$
$$u(t+h) = u(t) - hA(u(t))$$

The (recursive) solution to the difference equation is called Euler's explicit approximation, and is

$$\begin{aligned} u(h) &= u_0 - hA(u_0) \\ u(2h) &= u_1 - hA(u_1) \\ &= u_0 - hA(u_0) - hA(u_0 - hA(u_0)) \\ u(3h) &= u_0 - hA(u_0) - hA(u_0 - hA(u_0)) - hAu_0 - hA(u_0 - hA(u_0 - hA(u_0)) \\ &\vdots \end{aligned}$$

Hmmm. If A is unbounded ($-\Delta$, for example), then there is no guarantee that $u(kh)$ exists, and the numerical approximation fails.

We therefore start over, but this time we use Euler's implicit (or backward) approximation:

$$\frac{u(t+h)-u(t)}{h} + A(u(\underbrace{t+h}_{!})) = 0, \quad \text{so}$$

$$u(t+h) + hA(u(h+t)) = u(t) \quad \text{or}$$

$$u(t+h) = (I+hA)^{-1}u(t)$$

The solution to the difference equation is recursive solution is

$$u(kh) = (I+hA)^{-k}u_0$$

The operator $(I+hA)^{-1}$ is well-behaved — even if A is an unbounded operator — since A is monotone.

Comparison: In one dimension, when $Ax = rx$ is just multiplication by a scalar, the implicit and explicit approximations' results are familiar. Let $t = hk$ be a fixed time. Then the implicit approximation is

$$(1+hr)^{-k} = \frac{1}{\left(1+\frac{rt}{k}\right)^k} \to e^{-rt} \quad \text{as } k \to \infty,$$

while the explicit Euler approximation in this case is

$$(1-hr)^k = \left(1-\frac{rt}{k}\right)^k \to e^{-r(t)} \quad \text{as } k \to \infty.$$

The implicit and explicit approximations

$$u_h(t) = \frac{1}{\left(1+\frac{rt}{k}\right)^k} \qquad \text{implicit approximation to } e^{-rt}$$

$$v_h(t) = \left(1-\frac{rt}{k}\right)^k \qquad \text{explicit approximation to } e^{-rt}$$

have the same limit, but are qualitatively very different. For example:

1) $u_h(t) > 0$ on $[0,\infty)$ (as is the limit e^{-rt}), but $v_h(t)$ is negative for odd k and large enough t.
2) $u_h(t)$ is bounded on $[0,\infty)$, but $\lim_{t\to\infty} v_h(t) = (-1)^k\infty$.
3) For $t \geq 0$, the implicit approximation $u_h(t) \downarrow e^{-rt}$ as $h \downarrow 0$, but $v_h(t)$ does not converge monotonically to e^{-rt}. For an easy proof, use the integral test

$$\log\left(1+\frac{rt}{k}\right) = \int_1^{1+\frac{rt}{k}} \frac{1}{z}\,dz \leq \int_1^{1+\frac{rt}{k}} 1\,dz = \frac{rt}{k}.$$

Multiply by $-k$ (and reverse the inequality) and exponentiate to get

$$\left(1+\frac{rt}{k}\right)^{-k} \geq e^{-rt}.$$

The superiority of Euler's implicit approximation is demonstrated even in the case of a bounded operator in one dimension.

For $h > 0$, the operator

$$J_h = (I+hA)^{-1} \tag{J_h}$$

is called the **resolvent** of A. It helps to think

$$(I + hA)^{-1} = \frac{1}{I + hA} \quad \text{where } h \text{ and } A \text{ are } \geq 0.$$

Then it's easy to remember that, no matter how big A is, the resolvent is bounded when A is "positive" (monotone).

It's fun to work out the resolvent for subgradients on R. For example,

$$\begin{aligned}
x = (I + h\,\mathrm{sgn}(\cdot))^{-1} y \quad &\iff \\
y \in (I + h\,\mathrm{sgn}(\cdot)) x \quad &\iff \\
y \in \begin{cases} \{x + h\}, & x > 0 \\ [-h, h], & x = 0 \\ \{x - h\}, & x < 0 \end{cases} \quad &\iff \\
x = \begin{cases} y - h, & y > h \\ 0, & y \in [-h, h] \\ y + h, & y < -h \end{cases}
\end{aligned}$$

It is often easiest to draw the graph of $I + hA$ and then invert it. Please draw the graphs of the functions in the example to see how much easier the graphical technique is.

Notice that the resolvent $(I + h\,\mathrm{sgn}())^{-1}$ is defined everywhere because the graph of sgn() "fills the gap" at zero. This is an important property of subgradients, and it is interesting when a general monotone operator has it. If a monotone operator A has no proper extension to a monotone operator, then it is called **maximal** monotone.

Maximal monotone operators (on a Hilbert space) are exactly the monotone operators whose resolvents are defined everywhere. This is the deepest theorem we shall encounter, and we omit the proof[5]. The point is that Euler's implicit approximations exist for maximal monotone operators.

Yoshida Approximations: The resolvents $(I + hA)^{-1}$ must surely approach the identity I in some sense as $h \downarrow 0$. This observation suggests the approximation

$$\frac{A}{I + hA} \approx A.$$

To turn this suggestion into a definition, observe that the definition of the resolvent (J_h) implies

$$\frac{A}{I + hA} = \frac{1}{h}\left[I - \frac{I}{I + hA}\right] = \frac{1}{h}\left[I - J_h\right].$$

We therefore define the **Yoshida approximation** of A to be

$$A_h = \frac{1}{h}\left[I - J_h\right]. \tag{Y}$$

Think of A as a positive number (or, more precisely, "A is the operation of multiplying by a positive number"). Then the following results are obvious:

$$\begin{aligned}
\|J_h x_2 - J_h x_1\| &= \left\| \frac{I}{I + hA} x_2 - \frac{I}{I + hA} x_1 \right\| \\
&= \frac{1}{1 + hA} \|x_2 - x_1\| \\
&\leq \|x_2 - x_1\|
\end{aligned} \qquad \big(\|J_h\| \big)$$

[5] H. Brezis, "Operateurs Maximaux Monotones et semi-groupes de contractions dans les espaces de Hilbert", North-Holland, 1973, p. 23ff.

so J_h *is a contraction*,

$$A_h = \frac{A}{I + hA} = AJ_h \qquad \big(AJ_h\big)$$

so $A_h x$ is in the range of A, and

$$\begin{aligned} \|A_h x_2 - A_h x_1\| &= \left\| \frac{A}{I+hA} x_2 - \frac{A}{I+hA} x_1 \right\| \\ &= \frac{1}{h} \frac{hA}{1+hA} \|x_2 - x_1\| \qquad \big(\|A_h\|\big) \\ &\leq \frac{1}{h} \|x_2 - x_1\| \end{aligned}$$

so A_h *is Lipschitz continuous* with Lipschitz constant $\frac{1}{h}$.

All three of the results above are true for general monotone operators A, and their proofs are given in the appendix, but let us see how to use them first.

Optimal Rates[6] Since the Yoshida approximation A_h is a Lipschitz continuous operator, there is a unique solution to the initial value problem

$$\begin{cases} \dfrac{d}{dt} u_h(t) + A_h(u_h(t)) = 0 \\ u_h(0) = u_0 \end{cases} \qquad (IVP_h)$$

Our goal is to understand how quickly u_h converges to the solution u of (IVP) as $h \downarrow 0$. (Semi-group theorists denote the solution $u(t) = e^{-At}u_0$, and the approximations $u_h(t) = e^{-A_h t}u_0$.)

The initial value problems (IVP) and (IVP_h) imply

$$\begin{cases} \dfrac{d}{dt}(u(t) - u_h(t)) + A(u(t)) - A_h(u_h(t)) = 0 \\ u(0) - u_h(0) = 0 \end{cases}$$

Take the inner product of both sides of the top equation with $u - u_h$ and recognize the chain rule:

$$\begin{aligned} 0 &= \left\langle u(t) - u_h(t), \frac{d}{dt}(u(t) - u_h(t)) \right\rangle + \big\langle A(u(t)) - A_h(u_h(t)), u(t) - u_h(t) \big\rangle \\ &= \frac{d}{dt} \frac{1}{2} \|u(t) - u_h(t)\|^2 + \big\langle A(u(t)) - A_h(u_h(t)), u(t) - u_h(t) \big\rangle. \end{aligned}$$

Integrate both sides and use the initial value $u(0) = u_h(0)$:

$$0 = \frac{1}{2} \|u(T) - u_h(T)\|^2 + \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), u(t) - u_h(t) \big\rangle \, dt.$$

The integrand is almost positive — by equation $\big(AJ_h\big)$ and the monotonicity of A, it would be positive if the rightmost u_h were replaced by $J_j u_h$. Aiming for this integrand suggests the manipulation:

$$\begin{aligned} \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), u(t) - u_h(t) \big\rangle \, dt &= \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), u(t) - J_h u_h(t) \big\rangle \, dt \\ &\quad + \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), J_h u_h(t) - u_h(t) \big\rangle \, dt \\ &= \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), u(t) - J_h u_h(t) \big\rangle \, dt \\ &\quad - h \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), A_h u_h(t) \big\rangle \, dt \end{aligned}$$

[6] Jim Rulla, "Error analysis for implicit approximations to solutions to Cauchy problems", *SIAM J. Num. Anal*, 33 (1996), pp. 68–87.

Expand the rightmost integrand using the identity

$$\begin{aligned}\langle a-b, b\rangle &= \frac{1}{2}\langle a-b, a+b\rangle - \frac{1}{2}\langle a-b, a-b\rangle \\ &= \frac{1}{2}\left(\|a\|^2 - \|b\|^2\right) - \frac{1}{2}\|a-b\|^2\end{aligned}$$

and collect the positive terms on the left side of the equation:

$$\begin{aligned}&\frac{1}{2}\|u(T)-u_h(T)\|^2 + \int_0^T \langle A(u(t)) - A_h(u_h(t)), u(t) - J_h u_h(t)\rangle + \frac{h}{2}\|A(u(t)) - A_h(u_h(t))\|^2\, dt \\ &\qquad = \frac{h}{2}\int_0^T \|A(u(t))\|^2 - \|A_h(u_h(t))\|^2\, dt\end{aligned}$$

This identity contains much information. On the left side are three non-negative terms, measuring the difference between the solution $u(t)$ and the approximation $u_h(t)$ in three different ways. In particular, the first term alone satisfies

$$\|u(T)-u_h(T)\|^2 \leq h\int_0^T \|A(u(t))\|^2 - \|A_h(u_h(t))\|^2\, dt \qquad (R_h)$$

In English: The rate of convergence of u_h to u in the Hilbert space X is completely determined by the rate of convergence of $A_h u_h$ to Au in the Hilbert space $L^2(0,T;X)$. Furthermore, we have

$$\int_0^T \|A_h(u_h(t))\|^2\, dt \leq \int_0^T \|A(u(t))\|^2\, dt$$

— reminiscent of Bessel's inequality. Consequently, we have the very coarse estimate

$$\|u(T)-u_h(T)\| \leq \sqrt{h}\sqrt{\int_0^T \|A(u(t))\|^2\, dt} = O(\sqrt{h}).$$

The rate $O(\sqrt{h})$ is **sublinear** — the power of h is less than 1. The wave equation is an example of an (IVP) whose convergence rates are arbitrarily close to $= O(\sqrt{h})$. In other words, sublinear convergence is the best result we can hope for in general.

We never observe *sub*linear rates of convergence for the heat equation, however. When (IVP) is the heat equation, we see

$$\|u(T)-u_h(T)\| = O(h)$$

— linear rates of convergence. Why? The key is in the integral.

$$\int_0^T \|A(u(t))\|^2\, dt.$$

Since $A(u(t)) = -\frac{d}{dt}u(t)$, we are free to investigate either of

$$\int_0^T \|A(u(t))\|^2\, dt = \int_0^T \left\|\frac{d}{dt}u(t)\right\|^2 dt.$$

The right choice, however, is to investigate the mixed expression

$$\int_0^T \|A(u(t))\|^2\, dt = -\int_0^T \left\langle A(u(t)), \frac{d}{dt}u(t)\right\rangle dt.$$

It would be great if we could integrate the integrand explicitly. Can we? The integrand on the right looks like a chain rule, *provided A is the derivative of some function.* Do we know any such operators? Sure!

Theorem: If $A = \partial\varphi$, then the chain rule integrates the term of interest as

$$\begin{aligned}\int_0^T \|A(u(t))\|^2 \, dt &= -\int_0^T \left\langle \partial\varphi(u(t)), \frac{d}{dt}u(t) \right\rangle \, dt \\ &= -\int_0^T \frac{d}{dt}\varphi(u(t)) \, dt \\ &= \varphi(u(0)) - \varphi(u(T)).\end{aligned}$$

Why does this integrated expression guarantee linear rates of convergence? There is already an explicit factor of h in R_h, but in general

$$\|Au\|_{L^2(0,T;X)} - \|A_h u_h\|_{L^2(0,T;X)} \downarrow 0 \text{ arbitrarily slowly,}$$

causing sublinear convergence. The integrated terms, however, are, more or less,

$$\begin{aligned}\|Au\|_{L^2(0,T;X)} - \|A_h u_h\|_{L^2(0,T;X)} &= \big[\varphi(u(0)) - \varphi(u(T))\big] - \big[\varphi(u_h(0)) - \varphi(u_h(T))\big] \\ &\le \langle \partial\varphi(u_h(T)), u(t) - u_h(T)\rangle \quad \text{(more or less)} \\ &\le \|A(J_h(T))\| \, \|u(t) - u_h(T)\| \,,\end{aligned}$$

and the rightmost term is already $O(h)$ from estimate (R_h).

I find that the technical details in the "more or less" distract from the beauty of the argument, so if you are satisfied with the explanation, I advocate that you stop reading here. The remainder of the notes is devoted to explaining the "more ore less" above, and proving the estimates $\big(\|J_h\|\big)$, $\big(AJ_h\big)$, $\big(\|A_h\|\big)$, and $\big(\varphi_h\big)$.

We need one more piece of information: if A is a sub-gradient, then so is the Yoshida approximation A_h. In fact, if $A = \partial\varphi$, then

$$A_h = \partial\varphi_h$$

where

$$\varphi_h(x) = \frac{h}{2}\|A_h x\|^2 + \varphi(J_h x) \qquad \big(\varphi_h\big)$$

is a convex function. Again, we postpone the proof until we see what it buys us.

The argument for subgradients is

$$\begin{aligned}&\frac{1}{2}\|u(T) - u_h(T)\|^2 + \int_0^T \big\langle A(u(t)) - A_h(u_h(t)), u(t) - J_h u_h(t)\big\rangle + \frac{h}{2}\|A(u(t)) - A_h(u_h(t))\|^2 \, dt \\ &\qquad = \frac{h}{2}\int_0^T \|A(u(t))\|^2 - \|A_h(u_h(t))\|^2 \, dt \\ &\qquad = \frac{h}{2}\left[\varphi(u(0)) - \varphi(u(T)) - \varphi_h(u(0)) + \varphi_h(u_h(T))\right] \\ &\qquad = \frac{h}{2}\left[\varphi(u(0)) - \varphi_h(u(0)) + \varphi_h(u_h(T)) - \varphi(u(T))\right],\end{aligned}$$

By equation (φ_h), the definition of subgradient, and the definition (Y) of the Yoshida approximation, the initial values satisfy

$$\begin{aligned}\varphi(u(0)) - \varphi_h(u(0)) &\le \langle A(u(0)), u(0) - J_h(u(0))\rangle \\ &= h\,\langle A(u(0)), A_h(u(0))\rangle \,,\end{aligned}$$

picking up and extra factor of h.

Similarly, the final values satisfy

$$\begin{aligned}\varphi_h(u_h(T)) - \varphi(u(T)) &\leq \frac{h}{2}\|A_h u_h(T)\|^2 + \langle \partial\varphi(J_h u_h(T)), J_h u_h(T) - u(T)\rangle \\ &= \frac{h}{2}\|A_h u_h(T)\|^2 + \langle A(J_h u_h(T)), J_h u_h(T) - u_h(T)\rangle + \langle A(J_h u_h(T)), u_h(T) - u(T)\rangle \\ &= \frac{h}{2}\|A_h u_h(T)\|^2 - h\,\langle A(J_h u_h(T)), A_h u_h(T)\rangle + \langle A(J_h u_h(T)), u_h(T) - u(T)\rangle\end{aligned}$$

The first two terms on the right side have the extra factor of h that we seek. The last term succumbs to the inequality

$$\frac{h}{2}\langle a, b\rangle \leq \frac{1}{4}\|ha\|^2 + \frac{1}{4}\|b\|^2$$

which means

$$\begin{aligned}\frac{1}{2}\,\|u(T) - u_h(T)\|^2 \leq \frac{h^2}{2}\,\langle A(u(0)), A_h(u(0))\rangle + \frac{h^2}{4}\|A_h u_h(T)\|^2 - \frac{h^2}{2}\,\langle A(J_h u_h(T)), A_h u_h(T)\rangle \\ + \frac{h^2}{4}\|A(J_h u_h(T))\|^2 + \frac{1}{4}\|u_h(T) - u(T)\|^2\end{aligned}$$

Subtract the last term from both sides and complete the square to find the $O(h)$ estimate

$$\frac{1}{4}\,\|u(T) - u_h(T)\|^2 \leq \frac{h^2}{2}\,\langle A(u(0)), A_h(u(0))\rangle + \frac{h^2}{4}\|A_h u_h(T) - A(J_h u_h(T))\|^2$$

////

Appendix: For completeness, we prove below the 4 lemmas $\big(\|J_h\|\big)$, $\big(AJ_h\big)$, $\big(\|A_h\|\big)$, and $\big(\varphi_h\big)$ used in the proof of optimal rates of convergence. The proofs are representative of the techniques used in monotonicity arguments.

Suppose A is monotone (see definition (M)) and $h \geq 0$. Then

$$\langle (I + hA)x_2 - (I + hA)x_1, x_2 - x_1\rangle \geq \langle x_2 - x_1, x_2 - x_1\rangle = \|x_2 - x_1\|^2$$

for any x_i in the domain of A. Let $x_i = J_h(y_i)$ so that $(I + hA)x_i = y_i$. Then

$$\langle y_2 - y_1, J_h(y_2) - J_h(y_1)\rangle \geq \|J_h(y)_2 - J_h(y_1)\|^2,$$

so J_h bf is monotone, and the Cauchy-Schwarz inequality guarantees

$$\|y_2 - y_1\|\,\|J_h(y_2) - J_h(y_1)\| \geq \|J_h(y_2) - J_h(y_1)\|^2.$$

Consequently, J_h is a **contraction**:

$$\|J_h(y_2) - J_h(y_1)\| \leq \|y_2 - y_1\|\,,$$

and inequality $\big(\|J_h\|\big)$ is proven.

Suppose now that $J_h(y) = x$. Then

$$\begin{aligned} y &\in x + hA(x) \quad \text{so} \\ y - x &\in hA(x) \quad \text{so} \\ \frac{1}{h}(y - x) &\in A(x).\end{aligned}$$

Substitute $x = J_h(y)$ and use the definition of the Yoshida approximation $\big(Y\big)$:

$$\frac{1}{h}(y - J_h(y)) = A_h(y) \in A(J_h(y)).$$

This proves $\big(AJ_h\big)$. A corollary is that A_h is monotone:

$$\begin{aligned}\langle A_h(x_2) - A_h(x_1), x_2 - x_1\rangle &= \langle A(J_h(x_2)) - A(J_h(x_1)), x_2 - x_1\rangle \\ &= \langle A(J_h(x_2)) - A(J_h(x_1)), J_h(x_2) - J_h(x_1)\rangle \\ &\quad + \langle A(J_h(x_2)) - A(J_h(x_1)), (x_2 - J_h(x_2)) - (x_1 - J_h(x_1))\rangle \\ &= \langle A(J_h(x_2)) - A(J_h(x_1)), J_h(x_2) - J_h(x_1)\rangle \\ &\quad + h\left\|A_h(x_2)) - A_h(x_1)\right\|^2 \geq 0.\end{aligned}$$

Omit the term $\langle A(J_h(x_2)) - A(J_h(x_1)), J_h(x_2) - J_h(x_1)\rangle$ on the penultimate line in the previous paragraph to estimate

$$\langle A_h(x_2) - A_h(x_1), x_2 - x_1\rangle \geq h\left\|A_h(x_2)) - A_h(x_1)\right\|^2.$$

The Cauchy-Schwarz inequality implies

$$\left\|A_h(x_2) - A_h(x_1)\right\| \left\|x_2 - x_1\right\| \geq h\left\|A_h(x_2)) - A_h(x_1)\right\|^2,$$

from which $\big(\,\|A_h\|\,\big)$ follows:

$$\left\|A_h x_2 - A_h x_1\right\| \leq \frac{1}{h}\left\|x_2 - x_1\right\|.$$

Finally, consider the case $A = \partial\varphi$, and note that $I = \partial\frac{1}{2}\|\cdot\|^2$ is also a subgradient. Furthermore, for a given y, the linear functional

$$\varphi_y(z) = \langle y, z\rangle$$

is convex (since it is linear), with

$$\partial\varphi_y(\cdot) = \langle y, \cdot\rangle.$$

Consequently,

$$\begin{aligned}(I + hA)x \ni y \quad &\text{iff (since } A = \partial\varphi) \\ \partial\left(\frac{1}{2}\|x\|^2 + h\varphi(x) - \langle y, x\rangle\right) \ni 0 \quad &\text{iff (by completing the square)} \\ \partial\left(\frac{1}{2}\|x - y\|^2 + h\varphi(z) - \frac{1}{2}\|y\|^2\right) \ni 0 \quad &\text{iff} \\ \frac{1}{2}\|z - y\|^2 + h\varphi(z) - \frac{1}{2}\|y\|^2 \quad &\text{has a minimum at } z = x\end{aligned}$$

We therefore define

$$\varphi_h(y) = \min_z \frac{1}{2}\|z - y\|^2 + h\varphi(z).$$

Note: The minimum of a (lower semi-continuous) convex function that tends to ∞ with its argument always attains its minimum on a Hilbert space. In fact, if $\varphi(x) = M < \infty$ is any function value, then $\varphi^{-1}(M)$ is a closed, bounded, convex set. Consequently, $\varphi^{-1}(M)$ is weak* compact, so there is a weak* minimizer, which is the (strong) minimizer.

The minimizer is $x = J_h(y)$, so

$$\varphi_h(y) = \frac{1}{2}\|J_h(y) - y\|^2 + h\varphi(J_h(y)) = \frac{h^2}{2}\|A_h(y)\|^2 + h\varphi(J_h(y)).$$

This is the result of equation $\big(\varphi_h\big)$.