11 Smoothing

Goal

Moreau envelope, duality between smoothness and strong convexity, faster convergence for simple nonsmooth functions

Alert 11.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 11.2: Problem

We revisit the nonsmooth minimization problem:

$$\inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}). \tag{11.1}$$

We have seen that the subgradient algorithm achieves the *optimal* rate $O(t^{-1/2})$ with even matching constants. On the other hand, we have also learned an accelerated algorithm that converges at $O(t^{-2})$, provided that f is $\mathsf{L}^{[1]}$-smooth and convex. The central question we aim to address in this lecture is:

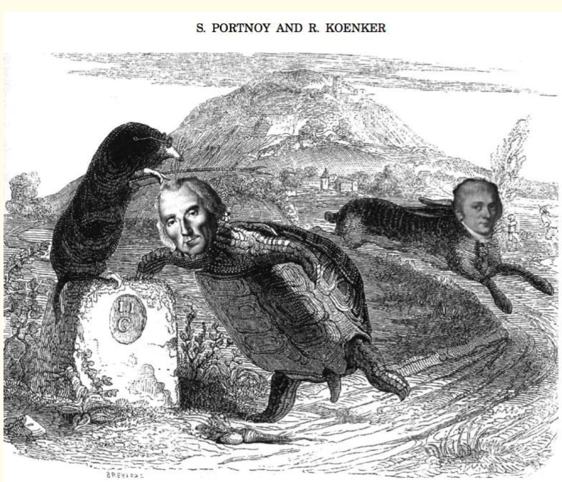
Can we do better, i.e. break the lower bound $O(t^{-1/2})$, at least for simple nonsmooth functions?

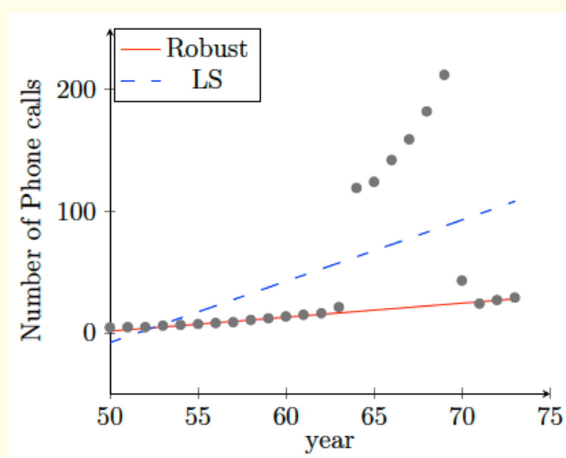
Example 11.3: Robust linear regression

Consider the following *robust* linear regression (with sparse regularization):

$$\min_{\mathbf{w}} \tfrac{1}{n}\|A\mathbf{w}-\mathbf{b}\|_1 + \lambda\|\mathbf{w}\|_1, \tag{11.2}$$

where we have replaced the least-square loss in linear regression with the least absolute loss, which has long been advocated since Laplace. The upside of the robust formulation above is that if there are some "grossly wrong" observations in A or $\mathbf{b}$, then the least absolute loss is less affected than the least-square loss. The downside is that the former is no longer $\mathsf{L}^{[1]}$-smooth for any $\mathsf{L}^{[1]} > 0$. Thus, robustness seems to incur a significant computational cost: $O(t^{-1/2})$ if we apply the subgradient algorithm, versus $O(t^{-2})$ for linear regression.





Alert 11.4: The familiar idea

Our main idea is relatively simple and familiar: We approximate a nonsmooth function with an $\mathsf{L}^{[1]}$-smooth one, just as in calculus where we approximate a smooth function by polynomials. In optimization we can only afford to find an approximate minimizer anyway, so a reasonable approximation of our objective function should not affect things much (intuitively).

However, since we do not know where the minimizer is, the approximation needs to be uniform (see next) and global (hence violating the black-box access assumption in our lower bound lecture).

Theorem 11.5: Uniform approximation leads to similar infimum

Consider the function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and its uniform approximation f_ϵ, i.e.,

$$\forall \mathbf{w}, \quad \underline{\epsilon} \le f(\mathbf{w}) - f_\epsilon(\mathbf{w}) \le \overline{\epsilon}.$$

Then, we have

$$\underline{\epsilon} \le \inf f - \inf f_\epsilon \le \overline{\epsilon}.$$

Moreover, let $f_\epsilon(\mathbf{w}) \le \inf f_\epsilon + \delta$, then $f(\mathbf{w}) \le \inf f + (\overline{\epsilon} - \underline{\epsilon}) + \delta$.

Proof: The first claim is immediate:

$$\inf_{\mathbf{w}} f(\mathbf{w}) - \inf_{\mathbf{w}} f_\epsilon(\mathbf{w}) \le \sup_{\mathbf{w}} [f(\mathbf{w}) - f_\epsilon(\mathbf{w})] \le \overline{\epsilon}.$$

The other direction can be proved similarly (or by switching the role of f and f_ϵ).

The last claim follows easily as well:

$$f(\mathbf{w}) \le f_\epsilon(\mathbf{w}) + \overline{\epsilon} \le \inf f_\epsilon + \overline{\epsilon} + \delta \le \inf f + (\overline{\epsilon} - \underline{\epsilon}) + \delta. \qquad \blacksquare$$

In other words, an δ-suboptimal minimizer $\mathbf{w}$ of the uniformly approximate function f_ϵ is $[(\overline{\epsilon} - \underline{\epsilon}) + \delta]$-suboptimal for the original function f. Our goal is then to:

- control the additional error $\overline{\epsilon} - \underline{\epsilon}$;
- choose f_ϵ with small $\mathsf{L}^{[1]}$-smoothness (if possible).

Note that we interpret $\infty - \infty = 0$ hence we must have $\mathrm{dom}\, f_\epsilon = \mathrm{dom}\, f$.

Example 11.6: Pointwise approximation is not enough

If for any $\mathbf{w}$, $f_\epsilon(\mathbf{w}) \to f(\mathbf{w})$ as $\epsilon \to 0$, then we say f_ϵ is a pointwise approximation of f. Clearly, uniform approximation implies pointwise approximation while the converse is not true, as the following example shows:

$$f_\epsilon(w) = \epsilon w,$$

which clearly converges to $f \equiv 0$ pointwise. However, $\inf f_\epsilon = -\infty < 0 = \inf f$ (thus uniform convergence fails).

Example 11.7: Close in function value implies close in minimizer?

In general we cannot say much about the closeness of the minimizers of f and its uniform approximation f_ϵ, no matter how small ϵ is. Indeed, fix $\epsilon > 0$ and pick a number t as large as you please. Then, consider the functions

$$f(w) = \begin{cases} \infty, & w < 0 \\ \frac{\epsilon}{t} w, & w \geq 0 \end{cases}, \quad f_\epsilon(w) = \begin{cases} \infty, & w < 0 \\ \epsilon - \frac{\epsilon}{t} w, & 0 \leq w \leq t \\ \frac{\epsilon}{t}(w - t), & w \geq t \end{cases}.$$

It is clear that f_ϵ is an ϵ-uniform approximation of f, but $\operatorname{argmin} f = 0$ while $\operatorname{argmin} f_\epsilon = t$.

Of course, the problem above is caused by the flatness of f. If f is of quadratic growth, then for any $\mathbf{w}$ such that $f_\epsilon(\mathbf{w}) \leq \inf f_\epsilon + \delta$, we have for any $\mathbf{w}_\star \in \operatorname{argmin} f$:

$$f(\mathbf{w}_\star) + \frac{\sigma}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2 \leq f(\mathbf{w}) \leq f(\mathbf{w}_\star) + \epsilon, \quad \text{where} \quad \epsilon = \overline{\epsilon} - \underline{\epsilon} + \delta.$$

Therefore, $\|\mathbf{w} - \mathbf{w}_\star\| \leq \sqrt{\frac{2\epsilon}{\sigma}}$, i.e. the approximate minimizer $\mathbf{w}$ (of f_ϵ) is necessarily close to its projection onto $\operatorname{argmin} f$.

Definition 11.8: Moreau envelope and infimal convolution (Moreau 1965)

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be closed and convex, then its Moreau envelope is defined as:

$$\mathrm{M}_f^\eta(\mathbf{z}) = \min_{\mathbf{w}} \frac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2 + f(\mathbf{w}), \quad \text{where} \quad \eta > 0,$$

while recall that the proximal map $\mathrm{P}_f^\eta(\mathbf{z})$ is the minimizer. Note that we always have

$$\forall \eta > 0, \quad \mathrm{M}_f^\eta \leq f$$

i.e. the envelope is always an under-approximation (e.g. $\underline{\epsilon} = 0$).

More generally, we define the infimal convolution of two functions f and g as:

$$(f \square g)(\mathbf{z}) := \Big[\inf_{\mathbf{w}} f(\mathbf{w}) + g(\mathbf{z} - \mathbf{w})\Big] = \Big[\inf_{\mathbf{w}} f(\mathbf{z} - \mathbf{w}) + g(\mathbf{w})\Big] = g \square f(\mathbf{z}),$$

which simply replaces the integral in the usual convolution with infimum. It is then clear that the Moreau envelope is the infimal convolution of f and the special (dilated) quadratic function $(\mathsf{q}\eta)(\mathbf{w}) := \eta \mathsf{q}(\mathbf{w}/\eta)$, where $\mathsf{q}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$.

Moreau, J. J. (1965). "Proximité et Dualtité dans un Espace Hilbertien". *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299.

Exercise 11.9: Infimum convolution preserves convexity

Prove that

- $\operatorname{sepi}(f \square g) = \operatorname{sepi} f + \operatorname{sepi} g, \quad \text{where} \quad \operatorname{sepi} h := \{(\mathbf{w}, t) : h(\mathbf{w}) < t\}$;
- f, g convex $\implies f \square g$ convex.

Definition 11.10: Fenchel Conjugate

Recall that the Fenchel conjugate of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined as:

$$f^*(\mathbf{w}^*) = \sup_{\mathbf{w}} \langle \mathbf{w}; \mathbf{w}^* \rangle - f(\mathbf{w}),$$

which is always closed and convex (even when f is not). (Our notation also indicates that the conjugate function is defined over the dual space where the gradient of f is from.) The so-called Fenchel-Young inequality follows from the definition:

$$f(\mathbf{w}) + f^*(\mathbf{w}^*) \geq \langle \mathbf{w}; \mathbf{w}^* \rangle,$$

with equality iff $\mathbf{w}^* = \partial f(\mathbf{w})$.

It can be shown that the biconjugate $f^{**} := (f^*)^*$ is the largest closed convex function that is dominated by f:

$$f^{**} = \sup\{g : g \leq f, g \text{ is closed and convex}\}.$$

In particular, when f is itself closed and convex, we have the marvelous duality result:

$$f^{**} = f. \tag{11.3}$$

(Technically, we need to interpret f^{**} as a function defined over the same space as f.)

It is known that $\mathsf{q} = \frac{1}{2}\|\cdot\|_2^2$ is the only function that is self-conjugate, i.e. $\mathsf{q} = \mathsf{q}^*$. Also,

$$(f \square g)^* = f^* + g^*$$
$$(f\lambda)^* = \lambda f^*, \quad \lambda > 0, \quad \text{where} \quad f\lambda := \lambda f(\cdot/\lambda).$$

For (closed) convex functions, applying the biconjugation (11.3) we also have $(f+g)^* = f^* \square g^*$ and $(\lambda f)^* = f^*\lambda$ (assuming the latter are closed).

Theorem 11.11: Duality between L-smoothness and $\frac{1}{\mathsf{L}}$-strong convexity

A convex function f is $\mathsf{L} = \mathsf{L}^{[1]}$-smooth iff f^* is $\frac{1}{\mathsf{L}}$-strongly convex.

Proof: This was mentioned and in fact proved in Alert 6.25. ■

Corollary 11.12: Moreau envelope is smooth and convex (Moreau 1965)

The Moreau envelope of a closed convex function is convex and $\frac{1}{\eta}$-smooth.

Proof: By definition $\mathrm{M}_f^\eta = f \square (\mathsf{q}\eta)$. Thus, $(\mathrm{M}_f^\eta)^* = f^* + \eta\mathsf{q}$ is η-strongly convex. Apply Theorem 11.11. ■

Conveniently, we have

$$\nabla \mathrm{M}_f^\eta(\mathbf{z}) = \tfrac{1}{\eta}(\mathbf{z} - \mathrm{P}_f^\eta(\mathbf{z})).$$

Moreau, J. J. (1965). "Proximité et Dualtité dans un Espace Hilbertien". *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299.

Exercise 11.13: Calculus for $\mathsf{L}^{[1]}$-smoothness

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be 1-smooth. Prove the following:

- $\eta\varphi$ is η-smooth;
- $\varphi\eta$ is $\frac{1}{\eta}$-smooth;
- $\varphi + \psi$ is 2-smooth if ψ is also 1-smooth;
- Let f and φ both be convex. Then, $f \square \varphi$ is also 1-smooth. In other words, infimal convolution with a smooth function is still smooth (when convexity is present).

Theorem 11.14: Uniform Moreau approximation (Beck and Teboulle 2012)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $\mathsf{L} = \mathsf{L}^{[0]}$-Lipschitz continuous (w.r.t. the norm $\|\cdot\|_2$). Then,

$$\forall \eta > 0, \quad \underbrace{\mathrm{M}_f^\eta \le}_{\underline{\epsilon}=0} f \le \mathrm{M}_f^\eta + \underbrace{\eta \mathsf{L}^2/2}_{\overline{\epsilon}}.$$

Proof: Indeed, using Lipschitz continuity we have

$$f(\mathbf{z}) - \mathrm{M}_f^\eta(\mathbf{z}) = \left[\sup_{\mathbf{w}} f(\mathbf{z}) - f(\mathbf{w}) - \tfrac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2\right] \le \left[\sup_{\mathbf{w}} \mathsf{L}\|\mathbf{z} - \mathbf{w}\|_2 - \tfrac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2\right] \le \eta \mathsf{L}^2/2,$$

which is what we wanted. ■

Beck, A. and M. Teboulle (2012). "Smoothing and First Order Methods: A Unified Framework". *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 557–580.

Alert 11.15: Trade-off

We immediately observe a trade-off in using the Moreau envelope to approximate an $\mathsf{L} = \mathsf{L}^{[0]}$-Lipschitz continuous function:

- the approximation error $\eta \mathsf{L}^2/2$ is proportional to η;
- the $\mathsf{L}^{[1]}$-smoothness of the approximation (Moreau envelope) is inversely proportional to η.

Not surprisingly, we now choose η carefully to get a faster algorithm.

Algorithm 11.16: Algorithmic implication of infimal smoothing

Let f be some nonsmooth $\mathsf{L} = \mathsf{L}^{[0]}$-Lipschitz continuous function. Then, to find $\mathbf{w}$ so that $f(\mathbf{w}) \le \inf f + \epsilon$ (namely, an ϵ-approximate minimizer):

- we can simply find $\mathbf{w}$ so that $\mathrm{M}_f^\eta(\mathbf{w}) \le \mathrm{M}_f^\eta(\mathbf{w}_\star) + \delta$, where $\mathbf{w}_\star \in \operatorname{argmin} f$;
- by Theorem 11.5 and Theorem 11.14, we know $f(\mathbf{w}) \le \inf f + 0 + \eta \mathsf{L}^2/2 + \delta$;
- thus, with $\eta \mathsf{L}^2/2 + \delta \le \epsilon$, $\mathbf{w}$ does the job.

If we use the accelerated Line 7 to minimize M_f^η, which is *bona fide* $\frac{1}{\eta}$-smooth, then the number of iterations we need to find $\mathbf{w}$ is (see Theorem 10.8):

$$\frac{2L\|\mathbf{w}_\star - \mathbf{z}_1\|_2^2}{(t+1)^2} = \frac{2\|\mathbf{w}_\star - \mathbf{z}_1\|_2^2}{\eta(t+1)^2} \le \delta \iff t \ge T := \sqrt{\frac{2}{\eta\delta}} \cdot \|\mathbf{w}_\star - \mathbf{z}_1\|_2 - 1.$$

To find the optimal trade-off, we solve:

$$\max_{\eta \mathsf{L}^2/2 + \delta \le \epsilon} \eta\delta \implies \delta = \epsilon/2, \quad \eta = \epsilon/\mathsf{L}^2 \implies T := \frac{2\mathsf{L}\|\mathbf{w}_\star - \mathbf{z}_1\|_2}{\epsilon} - 1.$$

which is significantly faster than the subgradient algorithm, which converges after $\frac{\mathsf{L}^2\|\mathbf{w}_\star - \mathbf{w}_0\|_2^2}{\epsilon^2} - 1$ iterations. We have seemingly beaten the lower bound!

In the above we need to fix the desired accuracy $\epsilon > 0$ in advance, since the parameter η depends on it. Alternatively, we can choose $\eta_t = O(1/t)$, then the rate of convergence can be shown to be $\frac{\log t}{t}$, which is still significantly better than the rate $O(1/\sqrt{t})$ of the subgradient algorithm.

Alert 11.17: The caveat

Let us recall that *evaluating*, let alone minimizing, M_f^η amounts to solving the following minimization problem:

$$\min_{\mathbf{w}} \quad f(\mathbf{w}) + \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2,$$

which does not seem to be any easier than our original problem (11.1)! It seems we have managed to reduce an easier (original) problem to a harder one (involving the Moreau envelope)...

Exercise 11.18: More calculus for $\mathsf{L}^{[1]}$-smoothness

The following simple observations turn out to be our rescue!

- Smoothing under linear transformation: Let M_f^η be a $\frac{1}{\eta}$-smooth approximation of f, then $\mathrm{M}_f^\eta(A\mathbf{w}+\mathbf{b})$ is a $\frac{\|A\|_{\mathrm{sp}}^2}{\eta}$-smooth approximation of $f(A\mathbf{w}+\mathbf{b})$.
- Smoothing under summation: $\sum_i \alpha_i \mathrm{M}_{f_i}^\eta$ is a $\frac{1}{\eta}$-smooth approximation of $\sum_i \alpha_i f_i$, where $\boldsymbol{\alpha} \in \Delta$.

Alert 11.19: What is the catch?

Let us examine the function $g(\mathbf{w}) = f(A\mathbf{w}+\mathbf{b})$ a bit more closely, where f is $\mathsf{L} = \mathsf{L}^{[0]}$-Lipschitz continuous hence g is $\mathsf{L}\|A\|_{\mathrm{sp}}$-Lipschitz continuous. We have two choices:

- Smooth f and then smooth g through composition:

$$\mathrm{M}_f^\eta(A\mathbf{w}+\mathbf{b}) \text{ is } \tfrac{\|A\|_{\mathrm{sp}}^2}{\eta}\text{-smooth and } \ \mathrm{M}_f^\eta(A\mathbf{w}+\mathbf{b}) \le g(\mathbf{w}) \le \mathrm{M}_f^\eta(A\mathbf{w}+\mathbf{b}) + \eta\mathsf{L}^2/2.$$

- Smooth g directly:

$$\mathrm{M}_g^\eta(\mathbf{w}) \text{ is } \tfrac{1}{\eta}\text{-smooth and } \ \mathrm{M}_g^\eta(\mathbf{w}) \le g(\mathbf{w}) \le \mathrm{M}_g^\eta(\mathbf{w}) + \eta\mathsf{L}^2\|A\|_{\mathrm{sp}}^2/2.$$

Repeating the calculation in Algorithm 11.16 we end up with the same convergence rate. However, note that the former could be much easier to compute than the latter. This is where the structure (e.g. composition with a linear map) of a function can help, and it demonstrates the potential harmless of being blind in the black-box view of optimization.

Example 11.20: Robust linear regression revisited

We have seen that the Moreau envelope of the absolute value function

$$\mathrm{M}_{|\cdot|}^\eta(z) = \Big[\min_w \tfrac{1}{2\eta}|w-z|^2 + |w|\Big] = \begin{cases} |z| - \frac{\eta}{2}, & \text{if } |z| \ge \eta \\ \frac{z^2}{2\eta}, & \text{if } |z| \le \eta \end{cases},$$

whence follows $\mathrm{M}_{\|\cdot\|_1}^\eta(\mathbf{z}) = \sum_j \mathrm{M}_{|\cdot|}^\eta(z_j)$. Thus, we may approximate the robust linear regression formulation (11.2) as:

$$\min_{\mathbf{w}} \ \frac{1}{n}\sum_i \mathrm{M}_{|\cdot|}^\eta(\langle \mathbf{a}_{i:}, \mathbf{w}\rangle + b_i) + \lambda\|\mathbf{w}\|_1.$$

which can now be solved using the accelerated gradient Line 7.

Alert 11.21: The price of smoothing

We point out that smoothing is not a free operation, for it increases the $\mathsf{L}^{[1]}$-smoothness parameter. Thus, whenever possible one should try to avoid smoothing any function unnecessarily. For instance, we could have also smoothed the ℓ_1-norm regularizer in Example 11.20 to arrive at:

$$\min_{\mathbf{w}} \ \frac{1}{n}\sum_i \mathrm{M}^{\eta}_{|\cdot|}(\langle \mathbf{a}_{i:}, \mathbf{w}\rangle + b_i) + \lambda \sum_j \mathrm{M}^{\eta}_{|\cdot|}(w_j),$$

whose L-smoothness parameter is evidently larger than the one in Example 11.20, leading to a slower convergence.

Exercise 11.22: Support vector machines (SVM) revisited

Recall the soft-margin SVM:

$$\min_{\mathbf{w}} \frac{1}{n}\sum_i (1 - y_i\hat{y}_i)_+ + \lambda\|\mathbf{w}\|_2^2, \quad \text{where} \quad \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b.$$

Explain how to find an ϵ-minimizer in $O(1/\epsilon)$ iterations.

History 11.23: Smoothing in optimization

To my knowledge, the first explicit exploitation of smoothing in gradient-type algorithms is due to Goldstein (1964). Moreau (1965) first proved that his envelope function is smooth but he did not apply this result to optimization. In some sense Nesterov (2005) rediscovered the duality between smoothness and strong convexity, and largely popularized it in optimization. Beck and Teboulle (2012) further developed the idea of smoothing using infimal convolution (in the context of gradient algorithms).

Goldstein, A. A. (1964). "Minimizing Functionals on Hilbert space", pp. 159–165.

Moreau, J. J. (1965). "Proximité et Dualtité dans un Espace Hilbertien". *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299.

Nesterov, Y. E. (2005). "Smooth Minimization of Non-Smooth Functions". *Mathematical Programming, Series A*, vol. 103, pp. 127–152.

Beck, A. and M. Teboulle (2012). "Smoothing and First Order Methods: A Unified Framework". *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 557–580.

Definition 11.24: Smoothing a sublinear function through perspectives

Recall that the recession function f^∞ of a closed convex function f with $0 \in \operatorname{dom} f$ is defined as

$$f^\infty(\mathbf{w}) = \sup_{\eta>0} \eta[f(\mathbf{w}/\eta) - f(\mathbf{0})] = \sup_{\eta>0} \frac{f(\mathbf{w}/\eta) - f(\mathbf{0})}{\frac{1}{\eta}} = \lim_{\eta\downarrow 0} \eta[f(\mathbf{w}/\eta) - f(\mathbf{0})] = \lim_{\eta\downarrow 0} \eta f(\mathbf{w}/\eta).$$

By definition, recession functions are positively homogeneous (and convex hence sublinear).

Thus, choosing a small η we may expect $f\eta$ to approximate f^∞ well (Beck and Teboulle 2012). Indeed,

$$\underbrace{-\eta f(\mathbf{0})}_{\epsilon} \le f^\infty(\mathbf{w}) - (f\eta)(\mathbf{w})$$

If we further assume

$$\forall \eta > 0, \quad f\eta \ge f^\infty \quad \text{hence} \ \inf_{\eta>0} f\eta = f^\infty,$$

so that $f\eta$ is an over-approximation of f^∞ (unlike the Moreau envelope which is an under-approximation).

We note that a given sublinear function f^∞ can be the recession function of multiple different functions f.

Beck, A. and M. Teboulle (2012). "Smoothing and First Order Methods: A Unified Framework". *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 557–580.

Example 11.25: Smoothing the max function

Let $f(\mathbf{w}) = \max_j w_j$ be the max function. Its Moreau envelope is:

$$\left[\min_{\mathbf{w}} \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + \max_j w_j\right] = \left[\min_t \min_{\mathbf{w}\le t} \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + t\right] = \left[\min_t \tfrac{1}{2\eta}\|(\mathbf{z}-t)_+\|_2^2 + t\right].$$

W.l.o.g. let $z_1 \geq \cdots \geq z_d$, and let $z_{k+1} \leq t < z_k$, then

$$\left[\inf_{t\in[z_{k+1}, z_k)} \frac{1}{2\eta}\sum_{j=1}^k (z_j - t)^2 + t\right] =: a_k.$$

Finding the smallest a_k gives us the solution for t hence $\mathbf{w} = t \wedge \mathbf{z}$.

Alternatively, the log-sum-exp function $\mathbf{w} \mapsto \log\sum_j \exp(w_j)$ can also be used to approximate the max:

$$\eta \log\sum_j \exp(w_j/\eta) - \eta\log d \leq \max_j w_j \leq \eta\log\sum_j \exp(w_j/\eta).$$

We note that max is the recession function of log-sum-exp:

$$\left[\lim_{\eta\downarrow 0} \eta\log\sum_j \exp(w_j/\eta)\right] = \left[\inf_{\eta>0} \eta\log\sum_j \exp(w_i/\eta)\right] = \max_j w_j.$$

Exercise 11.26: Smoothing the absolute

Show that the absolute function $|w|$ is the recession function of $\sqrt{w^2+1}$, which is 1-smooth. This function was used by Goldstein (1964).

Goldstein, A. A. (1964). "Minimizing Functionals on Hilbert space", pp. 159–165.