CS794/CO673: Optimization for Data Science

Lec 11: Smoothing

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

October 21, 2022

Problem

Composite smooth minimization:

$$f_\star = \inf_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

- f: nonsmooth but convex
- Subgradient achieves *optimal* rate $O(t^{-1/2})$, even with matching constants!
- Nesterov's momentum enjoys faster rate $O(t^{-2})$, provided that f is L-smooth

Can we break the lower bound $O(t^{-1/2})$, at least for *some* nonsmooth functions?

Problem

Composite smooth minimization:

$$f_\star = \inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

- f: nonsmooth but convex
- Subgradient achieves *optimal* rate $O(t^{-1/2})$, even with matching constants!
- Nesterov's momentum enjoys faster rate $O(t^{-2})$, provided that f is L-smooth

Can we break the lower bound $O(t^{-1/2})$, at least for *some* nonsmooth functions?

Problem

Composite smooth minimization:

$$f_\star = \inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

- f: nonsmooth but convex
- Subgradient achieves *optimal* rate $O(t^{-1/2})$, even with matching constants!
- Nesterov's momentum enjoys faster rate $O(t^{-2})$, provided that f is L-smooth

Can we break the lower bound $O(t^{-1/2})$, at least for *some* nonsmooth functions?

Problem

Composite smooth minimization:

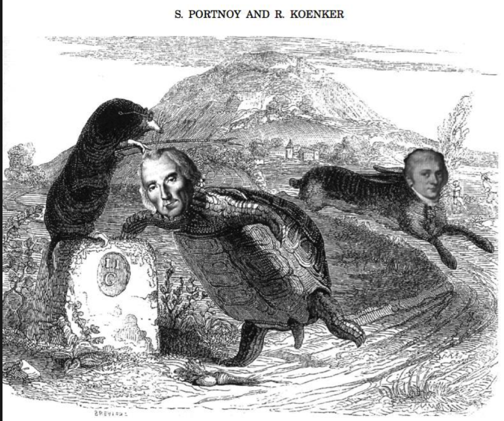
$$f_\star = \inf_{\mathbf{w} \in \mathbb{R}^d} \; f(\mathbf{w})$$

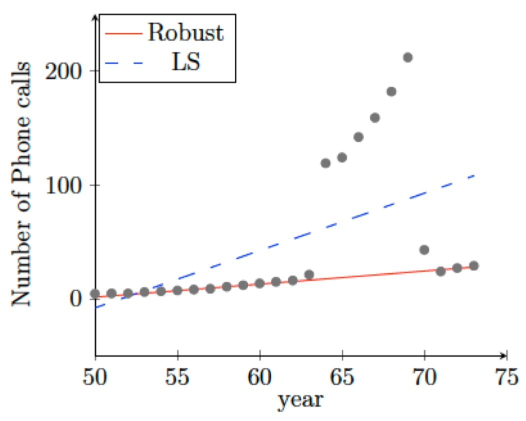
- f: nonsmooth but convex
- Subgradient achieves *optimal* rate $O(t^{-1/2})$, even with matching constants!
- Nesterov's momentum enjoys faster rate $O(t^{-2})$, provided that f is L-smooth

Can we break the lower bound $O(t^{-1/2})$, at least for *some* nonsmooth functions?

Robust Linear Regression

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





Approximation

- We approximate a nonsmooth function with an $L^{[1]}$-smooth one
  - just as in calculus where we approximate a smooth function by polynomials
- 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 lower bounds).

Approximation

- We approximate a nonsmooth function with an $L^{[1]}$-smooth one
  - just as in calculus where we approximate a smooth function by polynomials
- 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 lower bounds).

Approximation

- We approximate a nonsmooth function with an $L^{[1]}$-smooth one
  - just as in calculus where we approximate a smooth function by polynomials
- 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 lower bounds).

Approximation

- We approximate a nonsmooth function with an $L^{[1]}$-smooth one
  - just as in calculus where we approximate a smooth function by polynomials
- 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 lower bounds).

Approximation

- We approximate a nonsmooth function with an L[1]-smooth one
  - just as in calculus where we approximate a smooth function by polynomials
- 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 lower bounds).

Theorem: Uniform approximation leads to similar minimum

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

- δ-suboptimal minimizer $\mathbf{w}$ of the uniformly approximate function f_ϵ is $[(\overline{\epsilon} - \underline{\epsilon}) + \delta]$-suboptimal for the original function f
- Control the additional error $\overline{\epsilon} - \underline{\epsilon}$
- Choose f_ϵ with small $\mathsf{L}^{[1]}$-smoothness (if possible)

Theorem: Uniform approximation leads to similar minimum

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

- δ-suboptimal minimizer $\mathbf{w}$ of the uniformly approximate function f_ϵ is $[(\overline{\epsilon} - \underline{\epsilon}) + \delta]$-suboptimal for the original function f
- Control the additional error $\overline{\epsilon} - \underline{\epsilon}$
- Choose f_ϵ with small $\mathsf{L}^{[1]}$-smoothness (if possible)

Theorem: Uniform approximation leads to similar minimum

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

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

Then, we have

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

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

- δ-suboptimal minimizer $\mathbf{w}$ of the uniformly approximate function f_ϵ is $[(\overline{\epsilon} - \underline{\epsilon}) + \delta]$-suboptimal for the original function f
- Control the additional error $\overline{\epsilon} - \underline{\epsilon}$
- Choose f_ϵ with small $\mathsf{L}^{[1]}$-smoothness (if possible)

Theorem: Uniform approximation leads to similar minimum

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

- δ-suboptimal minimizer $\mathbf{w}$ of the uniformly approximate function f_ϵ is $[(\overline{\epsilon} - \underline{\epsilon}) + \delta]$-suboptimal for the original function f
- Control the additional error $\overline{\epsilon} - \underline{\epsilon}$
- Choose f_ϵ with small $\mathsf{L}^{[1]}$-smoothness (if possible)

Example: 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).

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \le f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\operatorname{argmin} \mathrm{M}_f^\eta = \operatorname{argmin} f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

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

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \le f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\operatorname{argmin} \mathrm{M}_f^\eta = \operatorname{argmin} f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

---

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

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \leq f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\operatorname{argmin} \mathrm{M}_f^\eta = \operatorname{argmin} f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

---

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

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \le f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\operatorname{argmin} \mathrm{M}_f^\eta = \operatorname{argmin} f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

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

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \ \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \le f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\operatorname{argmin} \mathrm{M}_f^\eta = \operatorname{argmin} f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

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

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \le f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\operatorname{argmin} \mathrm{M}_f^\eta = \operatorname{argmin} f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

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

Proximal Map and Moreau Envelope

$$\mathrm{P}_f^\eta(\mathbf{w}) := \underset{\mathbf{z}}{\mathrm{argmin}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$\mathrm{M}_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \ \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $\mathrm{P}_f^\eta : \mathbb{R}^d \to \mathbb{R}^d$ while $\mathrm{M}_f^\eta : \mathbb{R}^d \to \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f:
  - $\mathrm{M}_f^\eta \le f$, $\inf \mathrm{M}_f^\eta = \inf f$, $\mathrm{argmin}\, \mathrm{M}_f^\eta = \mathrm{argmin}\, f$
  - $\mathrm{M}_f^\eta \to f$ if $\eta \to 0$, and M_f^η is "smoother" than f

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

Fenchel Conjugate

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).

- Fenchel-Young inequality follows from the definition:

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

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

- $f^{**} = f$ iff f is (closed) convex

Fenchel Conjugate

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).

- Fenchel-Young inequality follows from the definition:

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

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

- $f^{**} = f$ iff f is (closed) convex

Fenchel Conjugate

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).

- Fenchel-Young inequality follows from the definition:

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

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

- $f^{**} = f$ iff f is (closed) convex

Theorem: 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.

Corollary:

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

$$(\mathrm{M}_f^\eta)^* = f^* + \eta \mathsf{q}$$

Example: Huber's function

$$h_\tau(s) := \begin{cases} \tau(|s| - \frac{\tau}{2}), & |s| \geq \tau \\ \frac{1}{2}s^2, & |s| \leq \tau \end{cases}$$

Theorem: Uniform Moreau approximation

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,

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

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

- 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

Theorem: Uniform Moreau approximation

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,

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

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

- 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

Theorem: Uniform Moreau approximation

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,

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

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

- 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

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

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

If we use Nesterov's momentum to minimize M_f^η:

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

To find the optimal trade-off, we solve:

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

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!

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

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

If we use Nesterov's momentum to minimize M_f^η:

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

To find the optimal trade-off, we solve:

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

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!

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

- can simply find $\mathbf{w}$ so that $\mathrm{M}_f^\eta(\mathbf{w}) \le \mathrm{M}_f^\eta(\mathbf{w}_\star) + \delta$, where $\mathbf{w}_\star \in \operatorname{argmin} f$;
- 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 Nesterov's momentum to minimize M_f^η:

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

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!

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

- can simply find $\mathbf{w}$ so that $\mathrm{M}_f^\eta(\mathbf{w}) \le \mathrm{M}_f^\eta(\mathbf{w}_\star) + \delta$, where $\mathbf{w}_\star \in \operatorname{argmin} f$;
- 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 Nesterov's momentum to minimize M_f^η:

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

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!

Example: Robust linear regression revisited

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

$$\mathrm{M}^{\eta}_{|\cdot|}(z) = \left[\min_{w} \frac{1}{2\eta}|w - z|^2 + |w|\right] = \begin{cases} |z| - \frac{\eta}{2}, & \text{if } |z| \geq \eta \\ \frac{z^2}{2\eta}, & \text{if } |z| \leq \eta \end{cases},$$

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

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

which can now be solved using Nesterov's momentum.

The Price of Smoothing

We point out that smoothing is not a free operation, for it increases the 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 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 the previous example, leading to a slower convergence.

Example: 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.

Example: 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 $z_1 \ge \cdots \ge z_d$, and let $z_{k+1} \le t < z_k$, 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 $\mathbf{w} = t \wedge \mathbf{z}$.

Example: Smoothing the max function, cont'

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:

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