Convex learning problems

CPSC 532S: Modern Statistical Learning Theory
14 February 2022
cs.ubc.ca/~dsuth/532S/22/

Admin

- In hybrid mode now:
  - Thursday office hours available both in-person (ICICS X563) and on Zoom

- A2 due Friday night
  - Groups of up to three, allowed separate per question
  - If you don’t have a group and want one, post on Piazza (asap)

- A1 grading: *allllllllmost* done – sorry

Practical learning algorithms

- So far, only really talked about ERM (and variants like SRM)

Practical learning algorithms

- So far, only really talked about ERM (and variants like SRM)
- Not always practical: e.g. NP-hard to maximize accuracy of a linear binary classifier

Practical learning algorithms

- So far, only really talked about ERM (and variants like SRM)
- Not always practical: e.g. NP-hard to maximize accuracy of a linear binary classifier
- A scheme that usually is practical: **convex** learning problems

Practical learning algorithms

- So far, only really talked about ERM (and variants like SRM)
- Not always practical: e.g. NP-hard to maximize accuracy of a linear binary classifier
- A scheme that usually is practical: **convex** learning problems
  - Can get an ε-approximate ERM with gradient descent:

Practical learning algorithms

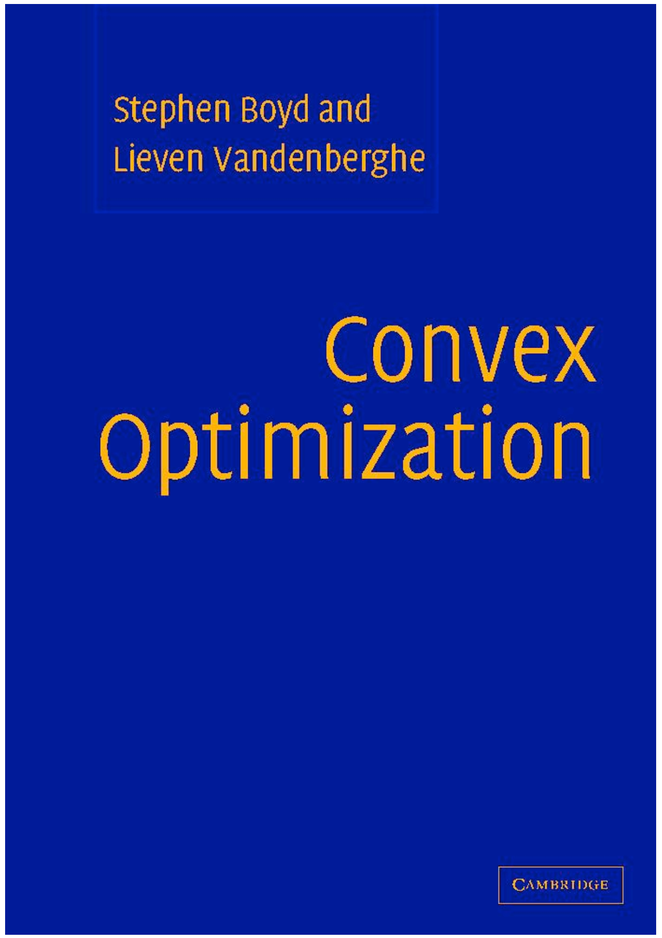
- So far, only really talked about ERM (and variants like SRM)
- Not always practical: e.g. NP-hard to maximize accuracy of a linear binary classifier
- A scheme that usually is practical: **convex** learning problems
  - Can get an ε-approximate ERM with gradient descent:
    - in $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps, if loss is convex and has Lipschitz gradients

Practical learning algorithms

- So far, only really talked about ERM (and variants like SRM)
- Not always practical: e.g. NP-hard to maximize accuracy of a linear binary classifier
- A scheme that usually is practical: **convex** learning problems
  - Can get an ε-approximate ERM with gradient descent:
    - in $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps, if loss is convex and has Lipschitz gradients
    - in $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$ steps, if loss is **strongly convex** with Lipschitz gradients

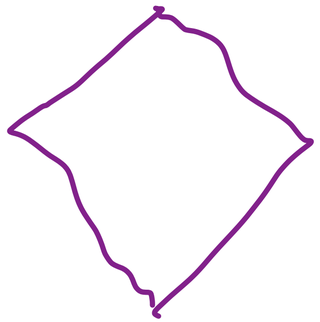
Convexity review

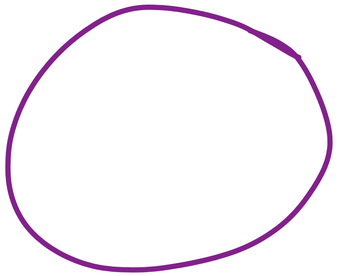
- Most of you probably already know most of this (based on the survey), but a quick reminder!
- For proofs and details, see e.g. chaps 2-3 of Boyd and Vandenberghe (free pdf)



Convex sets

- If $x, y \in C$, then $\alpha x + (1 - \alpha)y \in C$ for all $\alpha \in [0,1]$





Convex functions

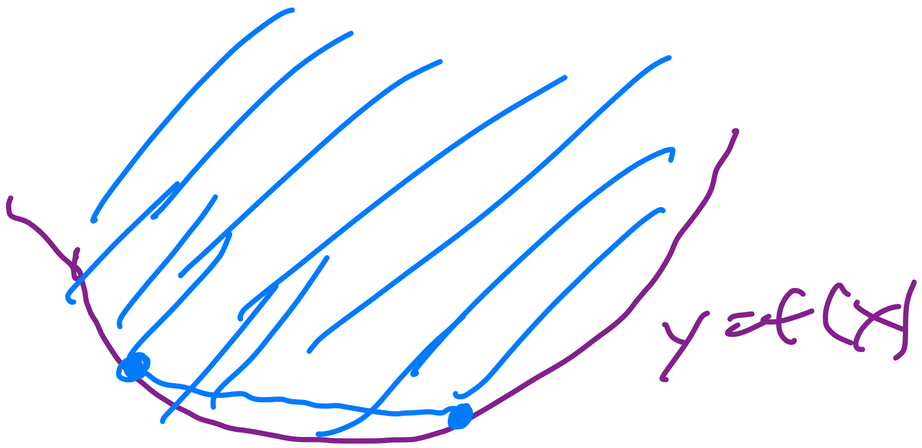
$f : \mathcal{X} \to \mathbb{R}$ is convex if, when $\mathcal{X}$ is a convex set:

- the **epigraph** $\left\{(x, r) \in \mathcal{X} \times \mathbb{R} : r \geq f(x)\right\}$ is a convex set

Convex functions

$f : \mathcal{X} \to \mathbb{R}$ is convex if, when $\mathcal{X}$ is a convex set:

- the **epigraph** $\{(x, r) \in \mathcal{X} \times \mathbb{R} : r \geq f(x)\}$ is a convex set
- it lies below its chords, $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ for $\alpha \in [0,1]$

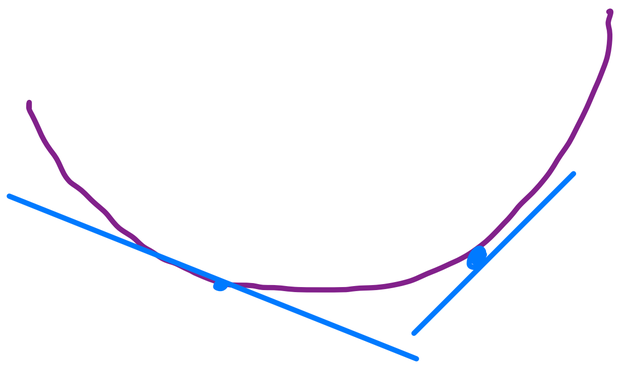


Convex functions

$f : \mathcal{X} \to \mathbb{R}$ is convex if, when $\mathcal{X}$ is a convex set:

- the **epigraph** $\{(x, r) \in \mathcal{X} \times \mathbb{R} : r \geq f(x)\}$ is a convex set
- it lies below its chords, $f\left(\alpha x + (1-\alpha)y\right) \leq \alpha f(x) + (1-\alpha)f(y)$ for $\alpha \in [0,1]$
- if f is differentiable: convex iff f lies above its tangent planes

$$f(x) \geq f(y) + [\nabla f(y)](x - y) \quad \text{for all } x, y$$



Convex functions

$f : \mathcal{X} \to \mathbb{R}$ is convex if, when $\mathcal{X}$ is a convex set:

- the **epigraph** $\{(x, r) \in \mathcal{X} \times \mathbb{R} : r \geq f(x)\}$ is a convex set
- it lies below its chords, $f\big(\alpha x + (1-\alpha)y\big) \geq \alpha f(x) + (1-\alpha) f(y)$ for $\alpha \in [0,1]$
- if f is differentiable: convex iff f lies above its tangent planes

$$f(x) \geq f(y) + [\nabla f(y)](x - y) \quad \text{for all } x, y$$

- if f is twice-differentiable: convex iff its Hessian $\nabla^2 f$ is positive semidefinite

$$\nabla^2 f \succeq 0 \quad \text{or} \quad \text{all eigenvalues} \geq 0 \quad \text{or} \quad v^\top [\nabla^2 f] v \geq 0 \text{ for all } v \quad \text{or} \quad \nabla^2 f = A^\top A$$

Operations that preserve convexity

If f and g are convex functions, then so are

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$
- $f + g$, or even $\int_{\mathcal{A}} f_y(x)\,\mathrm{d}w(y)$ if each f_y is convex and w a measure

$\ell(\cdot, y_i) = (\cdot - y_i)^2$

$L_S(\hat{y})$ $\hat{y} \in \mathbb{R}^n$

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$
- $f + g$, or even $\int_{\mathcal{A}} f_y(x)\,\mathrm{d}w(y)$ if each f_y is convex and w a measure
- $x \mapsto f(Ax + b)$

$$\ell_i(\hat{y}) = (\hat{y} - y_i)^2$$

$$\tilde{\ell}_i(w) = \ell_i(wx_i) = (wx - y_i)^2$$

$$L_S(w) \qquad L_D(w)$$

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$
- $f + g$, or even $\int_{\mathcal{A}} f_y(x)\,\mathrm{d}w(y)$ if each f_y is convex and w a measure
- $x \mapsto f(Ax + b)$
- $x \mapsto g(f(x))$ if $f : \mathcal{X} \to \mathbb{R}$ and g is nondecreasing

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$
- $f + g$, or even $\int_{\mathcal{A}} f_y(x)\,\mathrm{d}w(y)$ if each f_y is convex and w a measure
- $x \mapsto f(Ax + b)$
- $x \mapsto g(f(x))$ if $f : \mathcal{X} \to \mathbb{R}$ and g is nondecreasing
- $x \mapsto \max(f(x), g(x))$, or even $x \mapsto \sup_{y \in \mathcal{A}} f_y(x)$

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$
- $f + g$, or even $\int_{\mathcal{A}} f_y(x)\,\mathrm{d}w(y)$ if each f_y is convex and w a measure
- $x \mapsto f(Ax + b)$
- $x \mapsto g(f(x))$ if $f : \mathcal{X} \to \mathbb{R}$ and g is nondecreasing
- $x \mapsto \max(f(x), g(x))$, or even $x \mapsto \sup_{y \in \mathcal{A}} f_y(x)$
- If $f(x, y)$ is convex in (x, y), then $h(x) = \inf_{y \in C} f(x, y)$ for nonempty convex C

Operations that preserve convexity

If f and g are convex functions, then so are

- αf for any $\alpha \geq 0$
- $f + g$, or even $\displaystyle\int_{\mathcal{A}} f_y(x)\,\mathrm{d}w(y)$ if each f_y is convex and w a measure
- $x \mapsto f(Ax + b)$
- $x \mapsto g(f(x))$ if $f : \mathcal{X} \to \mathbb{R}$ and g is nondecreasing
- $x \mapsto \max(f(x), g(x))$, or even $x \mapsto \sup_{y \in \mathcal{A}} f_y(x)$
- If $f(x, y)$ is convex in (x, y), then $h(x) = \inf_{y \in C} f(x, y)$ for nonempty convex C
- Perspective transform: $h(x, t) = tf\left(\frac{x}{t}\right)$ for $t > 0$

(pause)

Convex learning problems

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ is **convex** if
  - $\mathscr{H}$ is a convex set
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex function

$\mathcal{H} = \{ x \mapsto w^\tau x : w \in \mathbb{R}^d \}$

$af + bg = x \mapsto af(x) + bg(x)$

$\mathcal{H}' = \{ w : w \in \mathbb{R}^d \}$

$\ell(h, z) = (\langle h, x \rangle - y)^2$

$h \in \mathbb{R}^d$, $z = (x, y)$

Convex learning problems

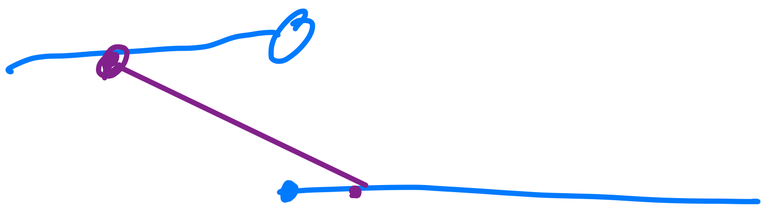
- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ is **convex** if
  - $\mathscr{H}$ is a convex set
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex function

- Example: linear regression with square loss

Convex learning problems

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ is **convex** if
  - $\mathscr{H}$ is a convex set
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex function

- Example: linear regression with square loss
- Non-example: linear classifiers with 0-1 loss



Convex learning problems

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ is **convex** if
  - $\mathscr{H}$ is a convex set
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex function

- Example: linear regression with square loss
- Non-example: linear classifiers with 0-1 loss

- For convex learning problems, ERM is a convex optimization problem

Convex learning problems

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ is **convex** if
  - $\mathscr{H}$ is a convex set
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex function

- Example: linear regression with square loss
- Non-example: linear classifiers with 0-1 loss

- For convex learning problems, ERM is a convex optimization problem
  - *Usually* implies learnable in polynomial time (but not always)

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w, (x, y)) = (wx - y)^2$

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w, (x, y)) = (wx - y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem

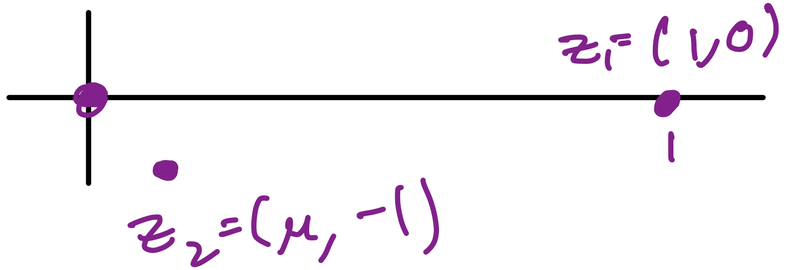
Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w, (x, y)) = (wx - y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w,(x,y)) = (wx-y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$
- Let $\mu = \frac{1}{2n}\log\frac{100}{99}$; $z_1 = (1,0)$ $z_2 = (\mu,-1)$; $\mathscr{D}_1(\{z\}) = \begin{cases} \mu & z = z_1 \\ 1-\mu & z = z_2 \end{cases}$; $\mathscr{D}_2(\{z_2\}) = 1$

≈ 0.01



Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w,(x,y)) = (wx-y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$
- Let $\mu = \frac{1}{2n}\log\frac{100}{99}$; $\begin{matrix} z_1 = (1,0) \\ z_2 = (\mu,-1) \end{matrix}$; $\mathscr{D}_1(\{z\}) = \begin{cases} \mu & z = z_1 \\ 1-\mu & z = z_2 \end{cases}$; $\mathscr{D}_2(\{z_2\}) = 1$
- At least 99% prob to only see z_2 in S: $\mathscr{D}_1^n\left((z_2,\ldots,z_2)\right) = (1-\mu)^n \geq e^{-2\mu n} = 0.99$

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w,(x,y)) = (wx-y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$
- Let $\mu = \frac{1}{2n}\log\frac{100}{99}$; $\begin{matrix} z_1 = (1,0) \\ z_2 = (\mu,-1) \end{matrix}$; $\mathscr{D}_1(\{z\}) = \begin{cases} \mu & z = z_1 \\ 1-\mu & z = z_2 \end{cases}$; $\mathscr{D}_2(\{z_2\}) = 1$
- At least 99% prob to only see z_2 in S: $\mathscr{D}_1^n\left((z_2,\ldots,z_2)\right) = (1-\mu)^n \geq e^{-2\mu n} = 0.99$
- Let $\hat{w} = A\left((z_2,\ldots,z_2)\right)$:

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w,(x,y)) = (wx-y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$
- Let $\mu = \frac{1}{2n}\log\frac{100}{99}$; $\begin{matrix} z_1 = (1,0) \\ z_2 = (\mu,-1) \end{matrix}$; $\mathscr{D}_1(\{z\}) = \begin{cases} \mu & z = z_1 \\ 1-\mu & z = z_2 \end{cases}$; $\mathscr{D}_2(\{z_2\}) = 1$
- At least 99% prob to only see z_2 in S: $\mathscr{D}_1^n\left((z_2,\ldots,z_2)\right) = (1-\mu)^n \geq e^{-2\mu n} = 0.99$
- Let $\hat{w} = A\left((z_2,\ldots,z_2)\right)$:
  - If $\hat{w} < \frac{-1}{2\mu}$, have $L_{\mathscr{D}_1}(\hat{w}) \geq \mu\cdot(\hat{w}-0)^2 \geq \frac{1}{4\mu} = 8n/\log\frac{100}{99} > 795n$.

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w,(x,y)) = (wx-y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$
- Let $\mu = \frac{1}{2n}\log\frac{100}{99}$; $\begin{matrix} z_1 = (1,0) \\ z_2 = (\mu,-1) \end{matrix}$; $\mathscr{D}_1(\{z\}) = \begin{cases} \mu & z = z_1 \\ 1-\mu & z = z_2 \end{cases}$; $\mathscr{D}_2(\{z_2\}) = 1$
- At least 99% prob to only see z_2 in S: $\mathscr{D}_1^n\left((z_2,\ldots,z_2)\right) = (1-\mu)^n \geq e^{-2\mu n} = 0.99$
- Let $\hat{w} = A\left((z_2,\ldots,z_2)\right)$:
  - If $\hat{w} < \frac{-1}{2\mu}$, have $L_{\mathscr{D}_1}(\hat{w}) \geq \mu\cdot(\hat{w}-0)^2 \geq \frac{1}{4\mu} = 8n/\log\frac{100}{99} > 795n$.
    - But $L_{\mathscr{D}_1}(0) = 1-\mu$, so excess error on $\mathscr{D}_1 \geq \frac{1}{4\mu} - 1 + \mu > 795n - 1 > \varepsilon$

Convex problems *aren't* necessarily learnable

- Consider (homogeneous) linear regression on $\mathbb{R}$, $\ell(w,(x,y)) = (wx-y)^2$
- Suppose a deterministic A can (agnostically) PAC learn this problem
- Take $\varepsilon = 0.01$, $\delta = 1/2$, n big enough that $L_{\mathscr{D}}(A(S)) - \inf_w L_{\mathscr{D}}(w) \leq \varepsilon$
- Let $\mu = \frac{1}{2n}\log\frac{100}{99}$; $\begin{matrix} z_1 = (1,0) \\ z_2 = (\mu,-1) \end{matrix}$; $\mathscr{D}_1(\{z\}) = \begin{cases} \mu & z = z_1 \\ 1-\mu & z = z_2 \end{cases}$; $\mathscr{D}_2(\{z_2\}) = 1$
- At least 99% prob to only see z_2 in S: $\mathscr{D}_1^n\left((z_2,\ldots,z_2)\right) = (1-\mu)^n \geq e^{-2\mu n} = 0.99$
- Let $\hat{w} = A\left((z_2,\ldots,z_2)\right)$:
  - If $\hat{w} < \frac{-1}{2\mu}$, have $L_{\mathscr{D}_1}(\hat{w}) \geq \mu\cdot(\hat{w}-0)^2 \geq \frac{1}{4\mu} = 8n/\log\frac{100}{99} > 795n.$
    - But $L_{\mathscr{D}_1}(0) = 1-\mu$, so excess error on $\mathscr{D}_1 \geq \frac{1}{4\mu} - 1 + \mu > 795n - 1 > \varepsilon$
  - If $\hat{w} \geq \frac{-1}{2\mu}$, then $L_{\mathscr{D}_2}(\hat{w}) \geq 1\cdot(\mu\hat{w}+1)^2 \geq (1-\frac{1}{2})^2 \geq \frac{1}{4}$, but $L_{\mathscr{D}_2}\left(\frac{-1}{\mu}\right) = 0$

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$
  - Optimal w on $\mathscr{D}_2$ is $-1/\mu \approx -200n$, which is really big

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$
  - Optimal w on $\mathscr{D}_2$ is $-1/\mu \approx -200n$, which is really big

- Counterexample for learning $\mathscr{H} = \{x \mapsto wx : |w| \leq 1\}$ with square loss:

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$
  - Optimal w on $\mathscr{D}_2$ is $-1/\mu \approx -200n$, which is really big

- Counterexample for learning $\mathscr{H} = \{x \mapsto wx : |w| \leq 1\}$ with square loss:
  - Exactly the same as before, but scale by $1/\mu$: $z_1 = (1/\mu, 0)$, $z_2 = (1, -1)$

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$
  - Optimal w on $\mathscr{D}_2$ is $-1/\mu \approx -200n$, which is really big

- Counterexample for learning $\mathscr{H} = \{x \mapsto wx : |w| \leq 1\}$ with square loss:
  - Exactly the same as before, but scale by $1/\mu$: $z_1 = (1/\mu, 0)$, $z_2 = (1, -1)$
  - wx on the old problem corresponds to $\frac{w}{\mu}(\mu x)$ here; same loss

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$
  - Optimal w on $\mathscr{D}_2$ is $-1/\mu \approx -200n$, which is really big

- Counterexample for learning $\mathscr{H} = \{x \mapsto wx : |w| \leq 1\}$ with square loss:
  - Exactly the same as before, but scale by $1/\mu$: $z_1 = (1/\mu, 0)$, $z_2 = (1, -1)$
  - wx on the old problem corresponds to $\frac{w}{\mu}(\mu x)$ here; same loss
  - Now $\mathscr{H}$ is bounded, but we have really big x values

Bounding $\mathscr{H}$ isn't enough

- Remember our bounds on $\mathfrak{R}_n$ for linear classes depended on bounding the norm of h, $\mathscr{H} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$
  - Optimal w on $\mathscr{D}_2$ is $-1/\mu \approx -200n$, which is really big

- Counterexample for learning $\mathscr{H} = \{x \mapsto wx : |w| \leq 1\}$ with square loss:
  - Exactly the same as before, but scale by $1/\mu$: $z_1 = (1/\mu, 0)$, $z_2 = (1, -1)$
  - wx on the old problem corresponds to $\frac{w}{\mu}(\mu x)$ here; same loss

  - Now $\mathscr{H}$ is bounded, but we have really big x values
    - oh, yeah…our bound on $\mathfrak{R}_n$ required bounding $\|x\|$ too!

Some learnable classes

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ **Convex ρ-Lipschitz B-Bounded** if
  - $\mathscr{H} \subset \mathbb{R}^d$ is a convex set, with $\|w\| \leq B$ for all $w \in \mathscr{H}$
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex, ρ-Lipschitz function

$$\ell_i(w) = |w x_i - y_i|$$

$$\ell_i(w) = (w x_i - y_i)^2$$

$$\ell_i' = 2(w x_i - y_i) \cdot x_i$$

Some learnable classes

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ **Convex ρ-Lipschitz B-Bounded** if
  - $\mathscr{H} \subset \mathbb{R}^d$ is a convex set, with $\|w\| \leq B$ for all $w \in \mathscr{H}$
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex, ρ-Lipschitz function

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ **Convex β-Smooth B-Bounded** if
  - $\mathscr{H} \subset \mathbb{R}^d$ is a convex set, with $\|w\| \leq B$ for all $w \in \mathscr{H}$
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex, nonnegative, β-smooth function
    - A function f is **β-smooth** if ∇f is β-Lipschitz

$\ell_i(w) = (w^T x_i - y_i)^2$

$\nabla \ell_i(w) = 2(w^T x_i - y_i) \cdot x_i$

$\|\nabla \ell_i(w) - \nabla \ell_i(v)\|$
$= \|2(w^T x_i - y_i) x_i - 2(w^T x_i - y_i) x_i\|$
$\leq 2|w^T x_i - y_i| \|x_i\|$

Some learnable classes

- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ **Convex ρ-Lipschitz B-Bounded** if
  - $\mathscr{H} \subset \mathbb{R}^d$ is a convex set, with $\|w\| \leq B$ for all $w \in \mathscr{H}$
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex, ρ-Lipschitz function

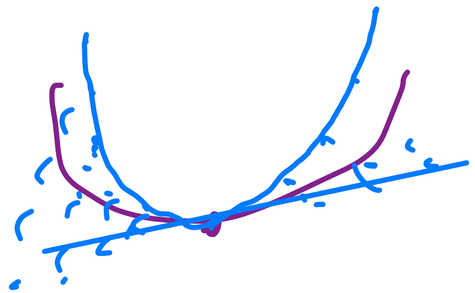
- A learning problem $(\mathscr{H}, \mathscr{Z}, \ell)$ **Convex β-Smooth B-Bounded** if
  - $\mathscr{H} \subset \mathbb{R}^d$ is a convex set, with $\|w\| \leq B$ for all $w \in \mathscr{H}$
  - for each $z \in \mathscr{Z}$, $\ell(\cdot, z)$ is a convex, nonnegative, β-smooth function
    - A function f is **β-smooth** if ∇f is β-Lipschitz

- We'll see soon that these classes are always learnable (efficiently!)

Key property of β-smooth f

$$f(y) - f(x) = \int_0^1 \nabla f\big(ty + (1-t)x\big) \cdot (y - x)\,\mathrm{d}t$$

$$= \int_0^1 \Big[\nabla f(x) + \nabla f\big(ty + (1-t)x\big) - \nabla f(x)\Big] \cdot (y - x)\,\mathrm{d}t$$



$$\leq \nabla f(x) \cdot (y - x) + \int_0^1 \Big\| \nabla f\big(ty + (1-t)x\big) - \nabla f(x) \Big\| \, \|y - x\|\,\mathrm{d}t$$

$$\leq \nabla f(x) \cdot (y - x) + \int_0^1 \beta \, \Big\| ty + (1-t)x - x \Big\| \, \|y - x\|\,\mathrm{d}t$$

$$= \nabla f(x) \cdot (y - x) + \beta \|y - x\|^2 \int_0^1 t\,\mathrm{d}t$$

$$= \nabla f(x) \cdot (y - x) + \frac{1}{2}\beta \|y - x\|^2$$

$A \succeq 0$ means A is psd

$A \succeq B$ means $A - B \succeq 0$

(implies $\nabla^2 f(y) \preceq \beta I$, if it exists)

(pause)

Gradient descent

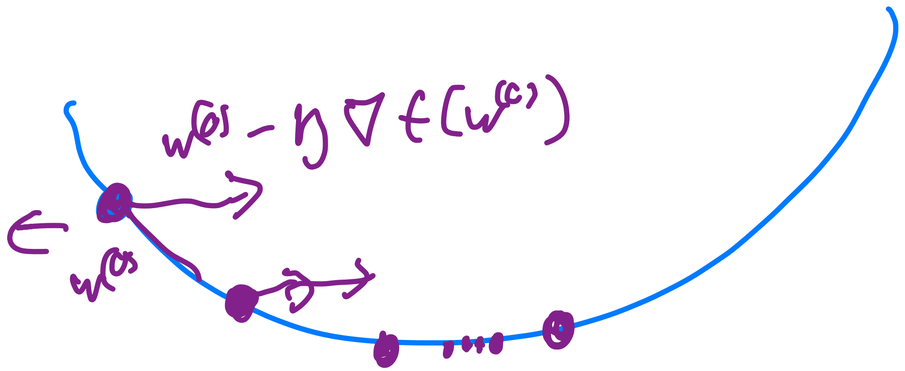
- Start at $w^{(0)}$, maybe 0 or sampled randomly. (SSBD calls this $w^{(1)}$)

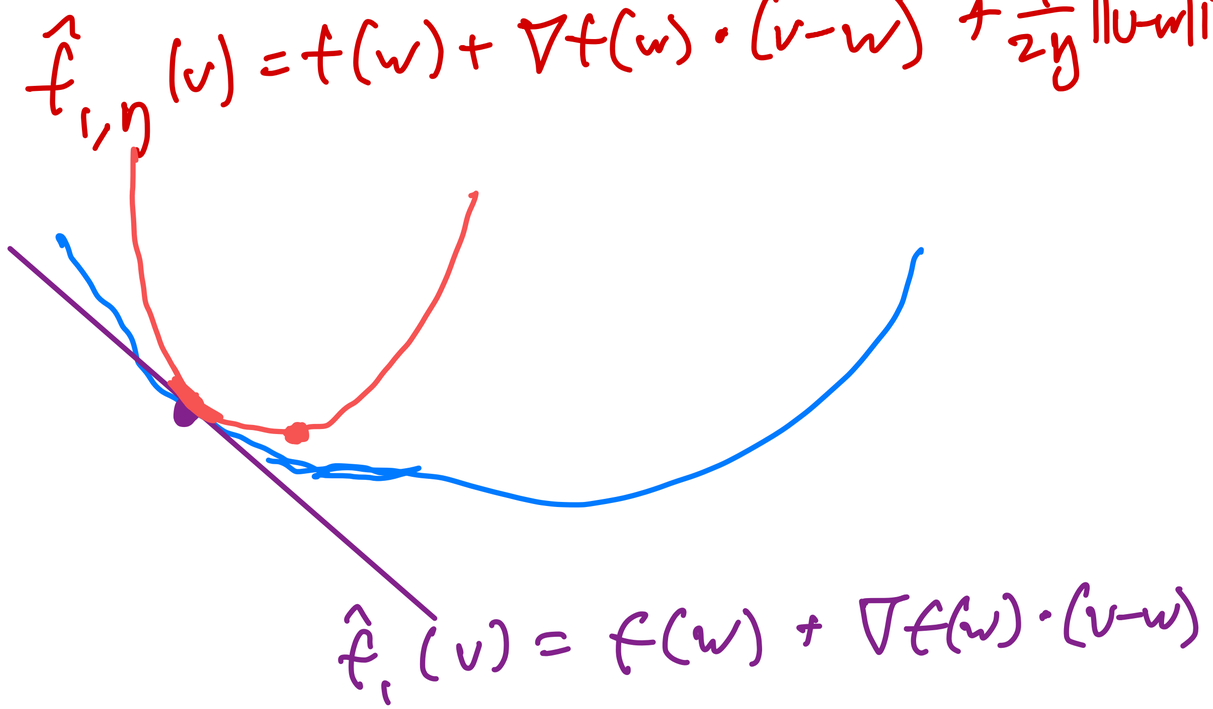
Gradient descent

- Start at $w^{(0)}$, maybe 0 or sampled randomly. (SSBD calls this $w^{(1)}$)
- Steps are $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$, where $\eta > 0$ is a **learning rate**

Gradient descent

- Start at $w^{(0)}$, maybe 0 or sampled randomly. (SSBD calls this $w^{(1)}$)
- Steps are $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$, where $\eta > 0$ is a **learning rate**
- Output last iterate $w^{(T)}$





Gradient descent

- Start at $w^{(0)}$, maybe 0 or sampled randomly. (SSBD calls this $w^{(1)}$)
- Steps are $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$, where $\eta > 0$ is a **learning rate**
- Output last iterate $w^{(T)}$
  - **Not** the only choice:

Gradient descent

- Start at $w^{(0)}$, maybe 0 or sampled randomly. (SSBD calls this $w^{(1)}$)
- Steps are $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$, where $\eta > 0$ is a **learning rate**
- Output last iterate $w^{(T)}$
  - **Not** the only choice:
    - SSBD use average of iterates, $\bar{w} = \frac{1}{T} \sum_{t=0}^{T-1} w^{(t)}$
    - Sometimes tail average $\frac{2}{T} \sum_{t=T/2}^{T} w^{(t)}$
    - Sometimes best iterate: $\operatorname{argmin}_{w^{(t)}:t\in[T]} f(w^{(t)})$
    - Or best on a validation set: $\operatorname{argmin}_{w^{(t)}:t\in[T]} L_V(w^{(t)})$

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

  very particular fixed η that *depends on length of optimization* T (also B, ρ)

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

  very particular fixed η that *depends on length of optimization* T (also B, ρ)

- We'll do something else (more standard):

  last iterate, β-smooth f, fixed $\eta \leq 1/\beta$

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

  very particular fixed η that *depends on length of optimization* T (also B, ρ)

- We'll do something else (more standard):

  last iterate, β-smooth f, fixed $\eta \leq 1/\beta$

  - We'll prove: $f(w^{(T)}) - f^* \leq \dfrac{\|w^{(0)} - w^*\|}{2\eta T}$

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

  very particular fixed η that *depends on length of optimization* T (also B, ρ)

- We'll do something else (more standard):

  last iterate, β-smooth f, fixed $\eta \leq 1/\beta$

  - We'll prove: $f(w^{(T)}) - f^* \leq \dfrac{\|w^{(0)} - w^*\|}{2\eta T}$
    - so can get suboptimality ε in $\mathcal{O}(1/\varepsilon)$ steps

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

  very particular fixed η that *depends on length of optimization* T (also B, ρ)

- We'll do something else (more standard):

  last iterate, β-smooth f, fixed $\eta \leq 1/\beta$

  - We'll prove: $f(w^{(T)}) - f^* \leq \dfrac{\|w^{(0)} - w^*\|}{2\eta T}$

    - so can get suboptimality ε in $\mathcal{O}(1/\varepsilon)$ steps

  - In practice, can be hard to compute β, and $1/\beta$ usually *too small*

$\|(w^{(k+1)} - w^{(k)})/\eta\|\nabla f(w^{(k)})\|$

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

  very particular fixed η that *depends on length of optimization* T (also B, ρ)

- We'll do something else (more standard):

  last iterate, β-smooth f, fixed $\eta \leq 1/\beta$

  - We'll prove: $f(w^{(T)}) - f^* \leq \dfrac{\|w^{(0)} - w^*\|}{2\eta T}$
    - so can get suboptimality ε in $\mathcal{O}(1/\varepsilon)$ steps

  - In practice, can be hard to compute β, and $1/\beta$ usually *too small*
  - **Backtracking line search** has a similar rate

Analysis of gradient descent

- SSBD section 14.1 analyzes:

  average iterate, initialize at 0, Lipschitz f

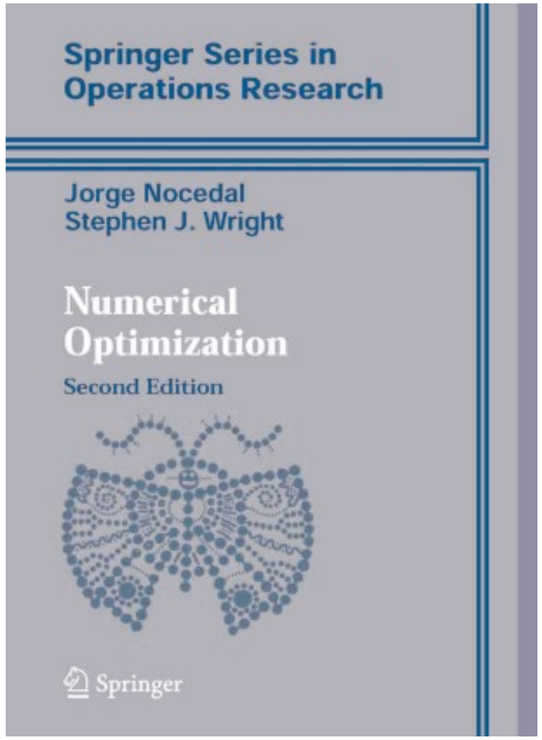
  very particular fixed η that *depends on length of optimization* T (also B, ρ)

- We'll do something else (more standard):

  last iterate, β-smooth f, fixed $\eta \leq 1/\beta$

  - We'll prove: $f(w^{(T)}) - f^* \leq \dfrac{\|w^{(0)} - w^*\|}{2\eta T}$
    - so can get suboptimality ε in $\mathcal{O}(1/\varepsilon)$ steps

  - In practice, can be hard to compute β, and $1/\beta$ usually *too small*
  - **Backtracking line search** has a similar rate



- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$f(w^+) \leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2$$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$\begin{aligned} f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\ &= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \end{aligned}$$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$\begin{aligned} f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\ &= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \\ &= f(w) - \eta \|\nabla f(w)\|^2 + \frac{\beta \eta^2}{2}\|\nabla f(w)\|^2 \end{aligned}$$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$
\begin{aligned}
f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\
&= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \\
&= f(w) - \eta \|\nabla f(w)\|^2 + \frac{\beta \eta^2}{2}\|\nabla f(w)\|^2 \\
&= f(w) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\nabla f(w)\|^2
\end{aligned}
$$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$
\begin{aligned}
f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\
&= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \\
&= f(w) - \eta \|\nabla f(w)\|^2 + \frac{\beta \eta^2}{2}\|\nabla f(w)\|^2 \\
&= f(w) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\nabla f(w)\|^2 \\
&\leq f(w) - \frac{\eta}{2}\|\nabla f(w)\|^2
\end{aligned}
$$

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$
\begin{aligned}
f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\
&= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \\
&= f(w) - \eta \|\nabla f(w)\|^2 + \frac{\beta \eta^2}{2}\|\nabla f(w)\|^2 \\
&= f(w) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\nabla f(w)\|^2 \\
&\leq f(w) - \frac{\eta}{2}\|\nabla f(w)\|^2
\end{aligned}
$$

"descent lemma": we're decreasing the objective! (note: didn't use convexity yet…)

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$
\begin{aligned}
f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\
&= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \\
&= f(w) - \eta \|\nabla f(w)\|^2 + \frac{\beta \eta^2}{2}\|\nabla f(w)\|^2 \\
&= f(w) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\nabla f(w)\|^2 \\
&\leq f(w) - \frac{\eta}{2}\|\nabla f(w)\|^2
\end{aligned}
$$

first-order convexity condition:

$$f(w^*) \geq f(w) + \nabla f(w)^\top (w^* - w)$$

so $f(w) \leq f(w^*) + \nabla f(w)^\top (w - w^*)$

"descent lemma": we're decreasing the objective! (note: didn't use convexity yet…)

- f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$
- β-smooth functions have $f(v) \leq f(w) + \nabla f(w)^\top (v - w) + \frac{\beta}{2}\|v - w\|^2$
- Iterate w goes to $w^+ = w - \eta \nabla f(w)$; plugging in to above, get

$$
\begin{aligned}
f(w^+) &\leq f(w) + \nabla f(w)^\top (w^+ - w) + \frac{\beta}{2}\|w^+ - w\|^2 \\
&= f(w) + \nabla f(w)^\top (-\eta \nabla f(w)) + \frac{\beta}{2}\|-\eta \nabla f(w)\|^2 \\
&= f(w) - \eta \|\nabla f(w)\|^2 + \frac{\beta \eta^2}{2}\|\nabla f(w)\|^2 \\
&= f(w) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\nabla f(w)\|^2 \\
&\leq f(w) - \frac{\eta}{2}\|\nabla f(w)\|^2
\end{aligned}
$$

first-order convexity condition:

$f(w^*) \geq f(w) + \nabla f(w)^\top (w^* - w)$

so $f(w) \leq f(w^*) + \nabla f(w)^\top (w - w^*)$

"descent lemma": we're decreasing the objective! (note: didn't use convexity yet…)

$$
\leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2}\|\nabla f(w)\|^2
$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w - w^*\|^2 + 2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w - w^*\|^2 + 2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \left[\|w - w^*\|^2 - 2\eta \nabla f(w)^\top (w - w^*) + \eta^2 \|\nabla f(w)\|^2\right]$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w - w^*\|^2 + 2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \left[\|w - w^*\|^2 - 2\eta \nabla f(w)^\top (w - w^*) + \eta^2 \|\nabla f(w)\|^2\right]$$

$$= \|w - w^*\|^2 - \|(w - w^*) - \eta \nabla f(w)\|^2$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w - w^*\|^2 + 2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \left[\|w - w^*\|^2 - 2\eta \nabla f(w)^\top (w - w^*) + \eta^2 \|\nabla f(w)\|^2\right]$$

$$= \|w - w^*\|^2 - \|(w - w^*) - \eta \nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w^+ - w^*\|^2$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w - w^*\|^2 + 2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \left[\|w - w^*\|^2 - 2\eta \nabla f(w)^\top (w - w^*) + \eta^2 \|\nabla f(w)\|^2\right]$$

$$= \|w - w^*\|^2 - \|(w - w^*) - \eta \nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w^+ - w^*\|^2$$

Plugging back in, $f(w^+) \leq f(w^*) + \frac{1}{2\eta} \left(\|w - w^*\|^2 - \|w^+ - w^*\|^2\right)$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

Have $f(w^+) \leq f(w^*) + \nabla f(w)^\top (w - w^*) - \frac{\eta}{2} \|\nabla f(w)\|^2$

$$2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w - w^*\|^2 + 2\eta \nabla f(w)^\top (w - w^*) - \eta^2 \|\nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \left[\|w - w^*\|^2 - 2\eta \nabla f(w)^\top (w - w^*) + \eta^2 \|\nabla f(w)\|^2\right]$$

$$= \|w - w^*\|^2 - \|(w - w^*) - \eta \nabla f(w)\|^2$$

$$= \|w - w^*\|^2 - \|w^+ - w^*\|^2$$

Plugging back in, $f(w^+) \leq f(w^*) + \frac{1}{2\eta} \left(\|w - w^*\|^2 - \|w^+ - w^*\|^2\right)$

and so $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta} \left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

$$f(w^{(T)}) - f(w^*) \leq \frac{1}{T}\sum_{k=1}^{T}\left(f(w^{(k)}) - f(w^*)\right)$$ (because of the descent lemma)

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

$$f(w^{(T)}) - f(w^*) \leq \frac{1}{T}\sum_{k=1}^{T}\left(f(w^{(k)}) - f(w^*)\right) \qquad \text{(because of the descent lemma)}$$

$$\leq \frac{1}{2\eta T}\sum_{k=1}^{T}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$$

$0 - 1 + 1 - 2 + 2 - \cdots + T - T-1$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

$$f(w^{(T)}) - f(w^*) \leq \frac{1}{T}\sum_{k=1}^{T}\left(f(w^{(k)}) - f(w^*)\right) \qquad \text{(because of the descent lemma)}$$

$$\leq \frac{1}{2\eta T}\sum_{k=1}^{T}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$$

$$= \frac{1}{2\eta T}\left(\|w^{(0)} - w^*\|^2 - \|w^{(T)} - w^*\|^2\right)$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

$$f(w^{(T)}) - f(w^*) \leq \frac{1}{T}\sum_{k=1}^{T}\left(f(w^{(k)}) - f(w^*)\right) \qquad \text{(because of the descent lemma)}$$

$$\leq \frac{1}{2\eta T}\sum_{k=1}^{T}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$$

$$= \frac{1}{2\eta T}\left(\|w^{(0)} - w^*\|^2 - \|w^{(T)} - w^*\|^2\right)$$

$$\leq \frac{1}{2\eta T}\|w^{(0)} - w^*\|^2$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

$$f(w^{(T)}) - f(w^*) \leq \frac{1}{T}\sum_{k=1}^{T}\left(f(w^{(k)}) - f(w^*)\right) \qquad \text{(because of the descent lemma)}$$

$$\leq \frac{1}{2\eta T}\sum_{k=1}^{T}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$$

$$= \frac{1}{2\eta T}\left(\|w^{(0)} - w^*\|^2 - \|w^{(T)} - w^*\|^2\right)$$

$$\leq \frac{1}{2\eta T}\|w^{(0)} - w^*\|^2 \qquad \leq \frac{B\beta}{2T} \text{ with } \eta = \frac{1}{\beta}$$

f is convex, ∇f is β-Lipschitz, learning rate $\eta \leq 1/\beta$; $w^+ = w - \eta \nabla f(w)$

now know $f(w^{(k)}) - f(w^*) \leq \frac{1}{2\eta}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$

$$f(w^{(T)}) - f(w^*) \leq \frac{1}{T}\sum_{k=1}^{T}\left(f(w^{(k)}) - f(w^*)\right) \qquad \text{(because of the descent lemma)}$$

$$\leq \frac{1}{2\eta T}\sum_{k=1}^{T}\left(\|w^{(k-1)} - w^*\|^2 - \|w^{(k)} - w^*\|^2\right)$$

$$= \frac{1}{2\eta T}\left(\|w^{(0)} - w^*\|^2 - \|w^{(T)} - w^*\|^2\right)$$

$$\leq \frac{1}{2\eta T}\|w^{(0)} - w^*\|^2 \qquad \leq \frac{B\beta}{2T} \text{ with } \eta = \frac{1}{\beta}$$

note $f\left(\frac{1}{T}\sum_{k=1}^{T} w^{(k)}\right) \leq \frac{1}{T}\sum_{k=1}^{T} f(w^{(k)})$: same for average iterate

So what?

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded classes
  - Will show they’re learnable; didn’t quite get there yet

So what?

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded classes
  - Will show they’re learnable; didn’t quite get there yet

So what?

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded classes
  - Will show they're learnable; didn't quite get there yet

- We can ~efficiently optimize empirical risk for Convex-Smooth-Bounded classes
  - SSBD section 14.1 shows (worse) rate for Convex-Lipschitz-Bounded, avg iterate
  - Turns out can get much faster rate if objective is **strongly convex**
  - Don't actually need differentiability everywhere: **subgradient descent**
  - For more: check out Bubeck (Convex Optimization: Algorithms and Complexity), Nocedal and Wright (Numerical Optimization), or take CPSC 536M

So what?

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded classes
  - Will show they're learnable; didn't quite get there yet

- We can ~efficiently optimize empirical risk for Convex-Smooth-Bounded classes
  - SSBD section 14.1 shows (worse) rate for Convex-Lipschitz-Bounded, avg iterate
  - Turns out can get much faster rate if objective is **strongly convex**
  - Don't actually need differentiability everywhere: **subgradient descent**
  - For more: check out Bubeck (Convex Optimization: Algorithms and Complexity), Nocedal and Wright (Numerical Optimization), or take CPSC 536M

- Next time:
  - How to use this / related stuff for learning guarantees
  - How to analyze classifiers (since 0-1 loss isn't Lipschitz or smooth)