SGD

CPSC 532S: Modern Statistical Learning Theory
16 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: still *alllllllllmost* done – sorry again

Last time: convex learning problems

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems
  - $\mathscr{H}$ is a convex set (e.g. convex set of parameters in $\mathbb{R}^d$), $||w|| \leq B$

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems
  - $\mathscr{H}$ is a convex set (e.g. convex set of parameters in $\mathbb{R}^d$), $\|w\| \leq B$
  - $\ell(\cdot, z)$ is ρ-Lipschitz or β-smooth

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems
  - $\mathscr{H}$ is a convex set (e.g. convex set of parameters in $\mathbb{R}^d$), $\|w\| \leq B$
  - $\ell(\cdot, z)$ is ρ-Lipschitz or β-smooth
    - β-smooth means that $\nabla \ell(\cdot, z)$ is β-Lipschitz

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems
  - $\mathscr{H}$ is a convex set (e.g. convex set of parameters in $\mathbb{R}^d$), $\|w\| \leq B$
  - $\ell(\cdot, z)$ is ρ-Lipschitz or β-smooth
    - β-smooth means that $\nabla \ell(\cdot, z)$ is β-Lipschitz
- Showed gradient descent can optimize convex β-smooth functions in $\frac{B\beta}{2\varepsilon}$ steps

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems
  - $\mathscr{H}$ is a convex set (e.g. convex set of parameters in $\mathbb{R}^d$), $\|w\| \leq B$
  - $\ell(\cdot, z)$ is ρ-Lipschitz or β-smooth
    - β-smooth means that $\nabla \ell(\cdot, z)$ is β-Lipschitz

- Showed gradient descent can optimize convex β-smooth functions in $\frac{B\beta}{2\varepsilon}$ steps
  - SSBD 14.1.1 shows $\frac{B^2\rho^2}{\varepsilon^2}$ steps for ρ-Lipschitz functions

Last time: convex learning problems

- Defined Convex-Lipschitz-Bounded and Convex-Smooth-Bounded problems
  - $\mathscr{H}$ is a convex set (e.g. convex set of parameters in $\mathbb{R}^d$), $\|w\| \leq B$
  - $\ell(\cdot, z)$ is ρ-Lipschitz or β-smooth
    - β-smooth means that $\nabla \ell(\cdot, z)$ is β-Lipschitz

- Showed gradient descent can optimize convex β-smooth functions in $\frac{B\beta}{2\varepsilon}$ steps
  - SSBD 14.1.1 shows $\frac{B^2\rho^2}{\varepsilon^2}$ steps for ρ-Lipschitz functions
- We can run ERM efficiently…but does it work statistically?

One thing I forgot to say…

- Convexity implies that any local minimum is a global minimum
  - We didn’t use this directly in the proof, but good to know!

One thing I forgot to say…

- Convexity implies that any local minimum is a global minimum
  - We didn't use this directly in the proof, but good to know!

- **Strict convexity** implies there's only one global minimum
  - $f\left(\alpha x_1 + (1-\alpha)x_2\right) > \alpha f(x_1) + (1-\alpha)f(x_2)$ for $\alpha \in (0,1)$
  - Hessian $\succ 0$ implies strictly convex, but converse not true (e.g. $f(x) = x^4$)

Convex surrogate losses

There's more to say, but just the basics for now:

Convex surrogate losses

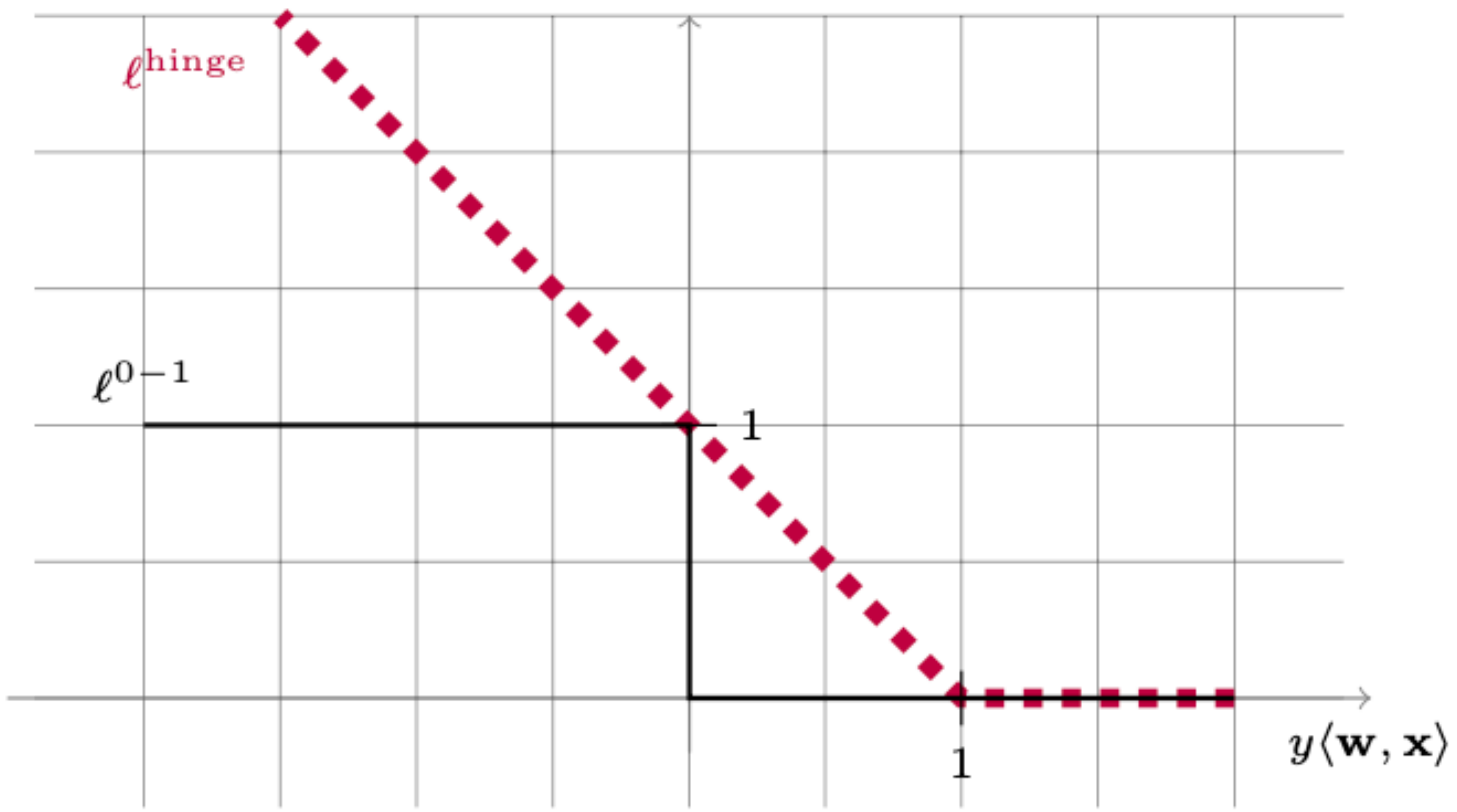
There's more to say, but just the basics for now:

- The 0-1 loss is not convex

Convex surrogate losses

There's more to say, but just the basics for now:

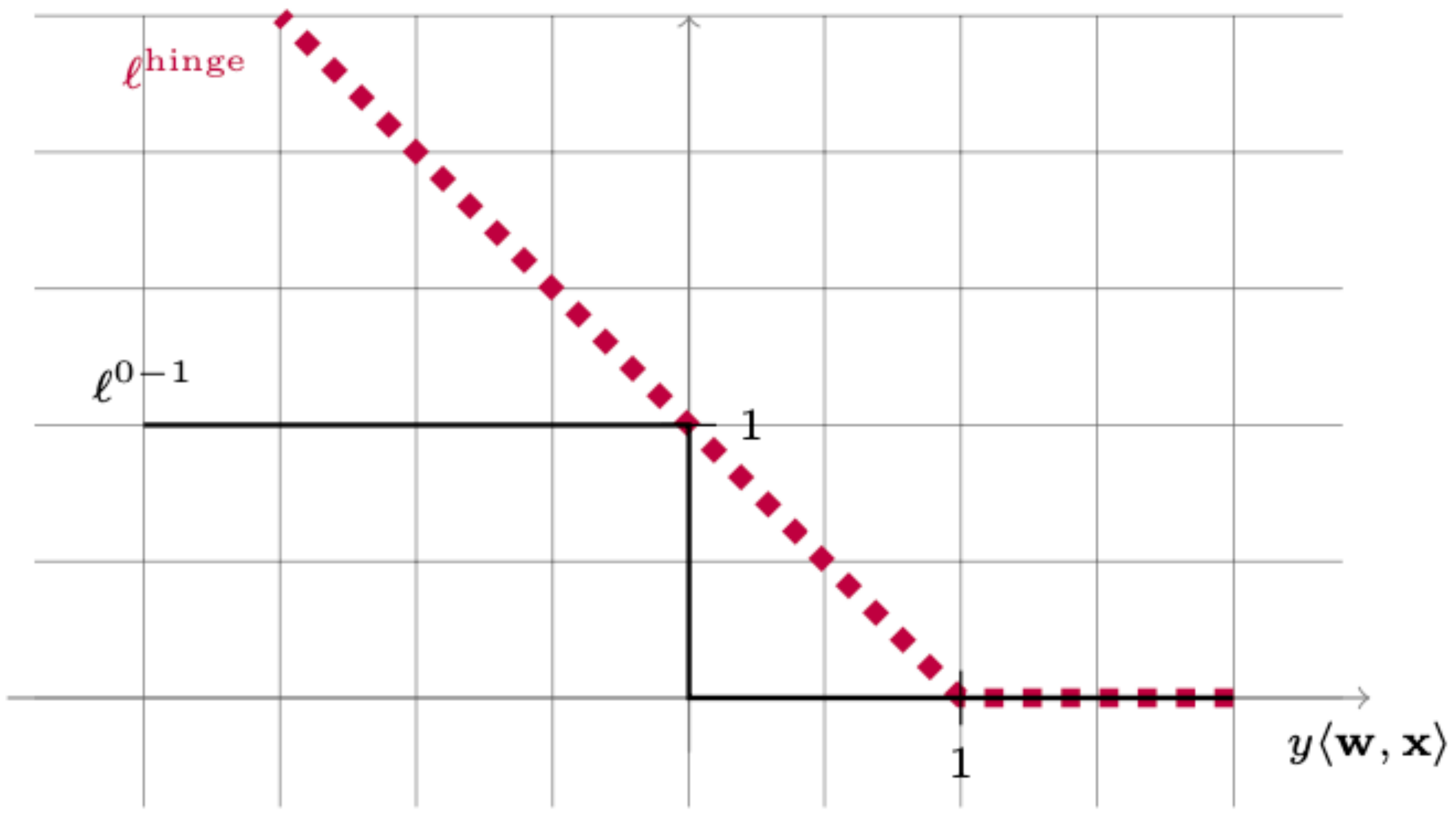
- The 0-1 loss is not convex
- *Hinge loss* is: $\ell(h, (x, y)) = \max\{0, 1 - yh(x)\}$



Convex surrogate losses

There's more to say, but just the basics for now:

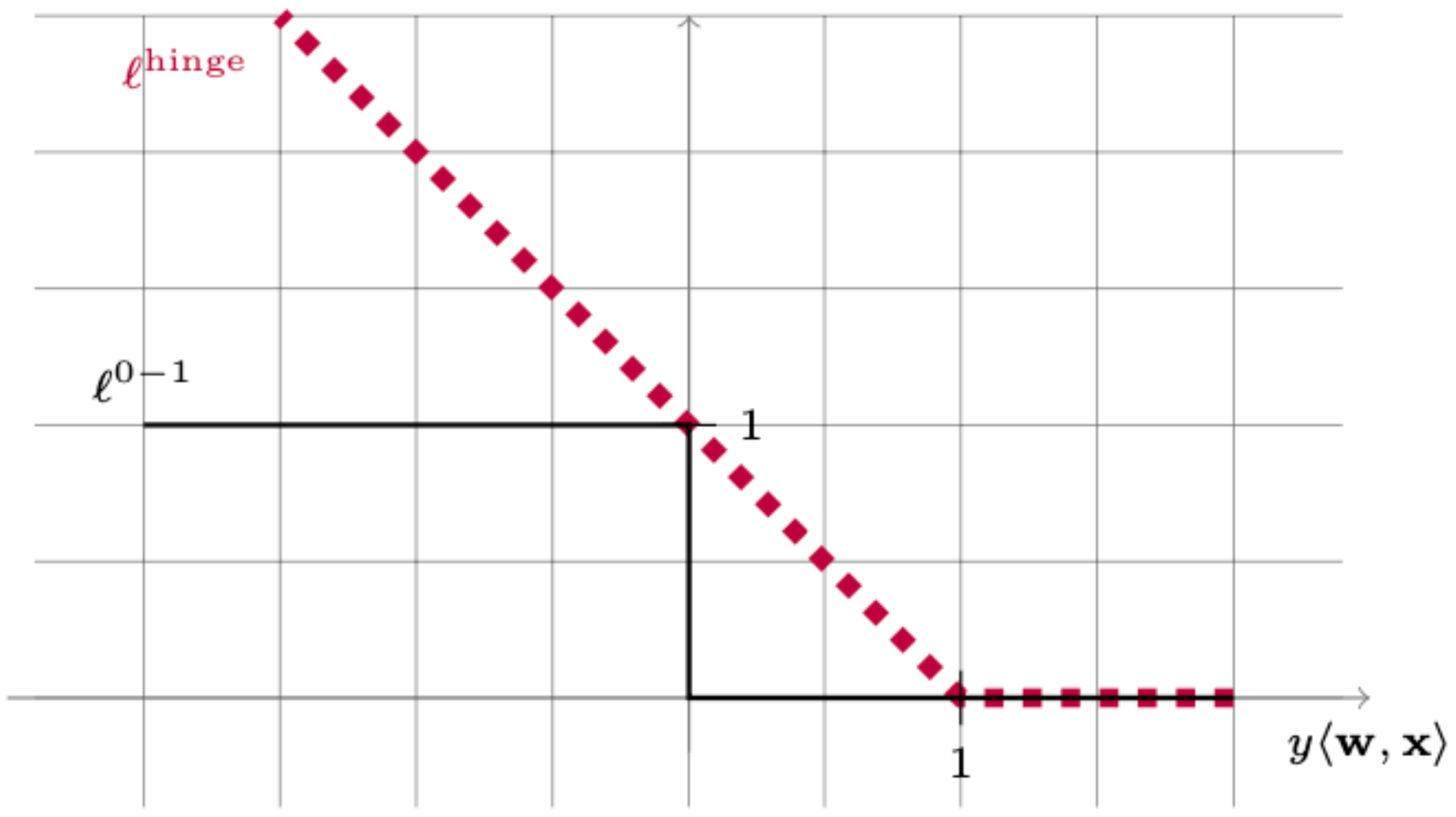
- The 0-1 loss is not convex
- *Hinge loss* is: $\ell(h,(x,y)) = \max\{0,1 - yh(x)\}$
- Also, $\ell^{0-1}(h,z) \leq \ell^{hinge}(h,z)$



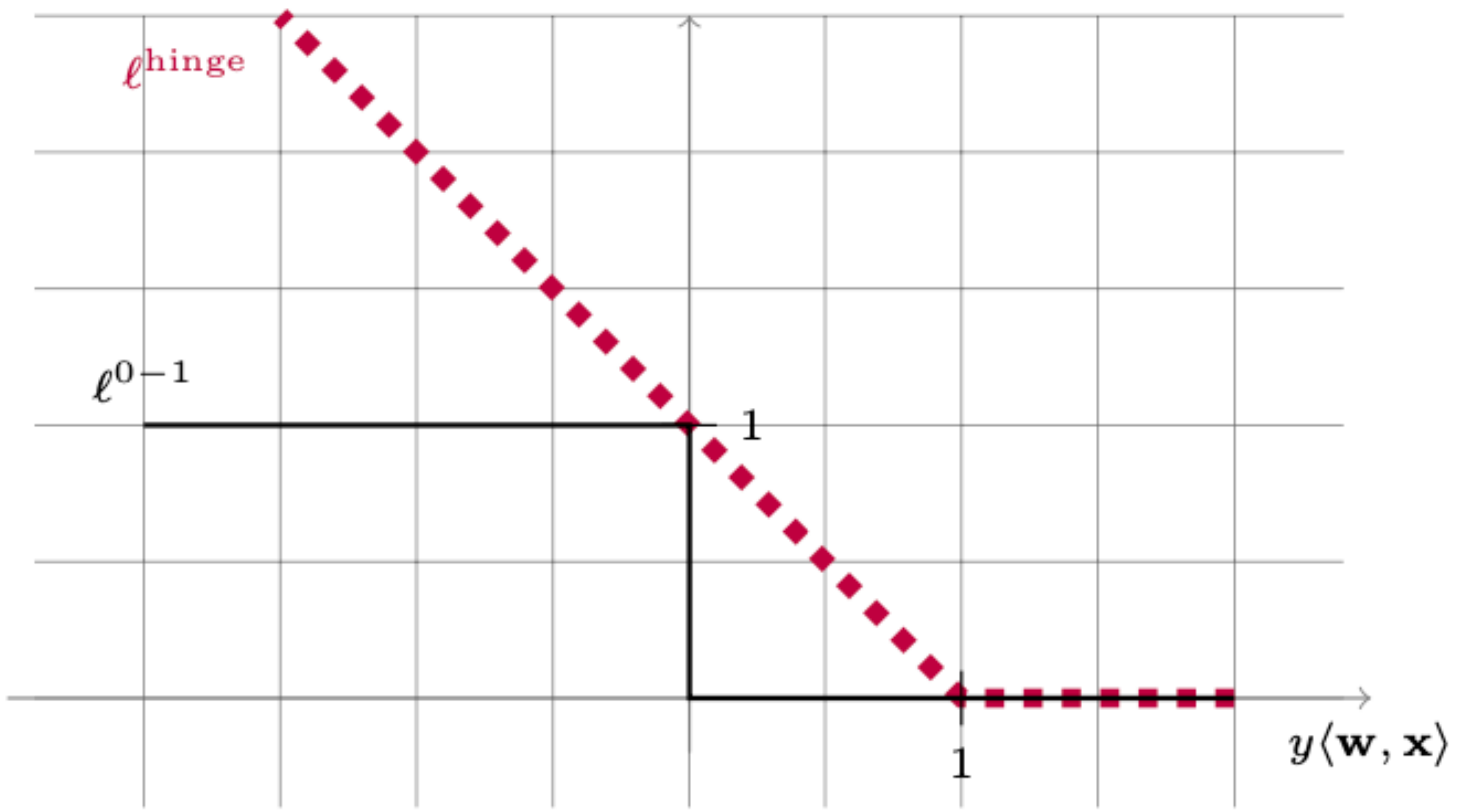
Convex surrogate losses

There's more to say, but just the basics for now:

- The 0-1 loss is not convex
- *Hinge loss* is: $\ell(h, (x, y)) = \max\{0, 1 - yh(x)\}$
- Also, $\ell^{0-1}(h, z) \leq \ell^{hinge}(h, z)$
  - So, $L_{\mathcal{D}}^{0-1}(h) \leq L_{\mathcal{D}}^{hinge}(h)$



Convex surrogate losses



There's more to say, but just the basics for now:

- The 0-1 loss is not convex
- *Hinge loss* is: $\ell(h, (x, y)) = \max\{0, 1 - yh(x)\}$
- Also, $\ell^{0-1}(h, z) \leq \ell^{hinge}(h, z)$
  - So, $L_{\mathcal{D}}^{0-1}(h) \leq L_{\mathcal{D}}^{hinge}(h)$
- $L_{\mathcal{D}}^{0-1}(\hat{h}) - L_{\mathcal{D}}^{0-1,*} \leq \left(\min_{h \in \mathcal{H}} L_{\mathcal{D}}^{0-1}(h) - L_{\mathcal{D}}^{0-1,*}\right) + \left(\min_{h \in \mathcal{H}} L_{\mathcal{D}}^{hinge}(h) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}^{0-1}(h)\right) + \left(L_{\mathcal{D}}^{hinge}(\hat{h}) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}^{hinge}(h)\right)$

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $\quad f(u) \geq f(w) + \langle u - w, v \rangle$

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $\quad f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the "differential set" for some reason)

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
    - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
    - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
        - (SSBD call this the “differential set” for some reason)

- If f is differentiable at w, the gradient is the only subderivative there

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $\quad f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the “differential set” for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $\quad f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the "differential set" for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one
- Subdifferential is always a nonempty, compact, convex set

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the "differential set" for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one
- Subdifferential is always a nonempty, compact, convex set
- A convex f is ρ-Lipschitz (on a convex open set) iff all of its subgradients have $\|v\| \leq \rho$

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the “differential set” for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one
- Subdifferential is always a nonempty, compact, convex set
- A convex f is ρ-Lipschitz (on a convex open set) iff all of its subgradients have $\|v\| \leq \rho$

- **Subgradient descent**: instead of a gradient, pick any subgradient

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the "differential set" for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one
- Subdifferential is always a nonempty, compact, convex set
- A convex f is ρ-Lipschitz (on a convex open set) iff all of its subgradients have $\|v\| \leq \rho$

- **Subgradient descent**: instead of a gradient, pick any subgradient
  - The analysis is exactly the same

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the “differential set” for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one
- Subdifferential is always a nonempty, compact, convex set
- A convex f is ρ-Lipschitz (on a convex open set) iff all of its subgradients have $\|v\| \leq \rho$

- **Subgradient descent**: instead of a gradient, pick any subgradient
  - The analysis is exactly the same
  - (kind of) what PyTorch/etc do for ReLU functions anyway!

Subgradients

- A **subgradient** of f at w is a vector v such that the tangent with normal v lies below f:
  - For all u in the domain of f, $f(u) \geq f(w) + \langle u - w, v \rangle$
  - The **subdifferential** at w, $\partial f(w)$, is the set of all valid subgradients
    - (SSBD call this the "differential set" for some reason)

- If f is differentiable at w, the gradient is the only subderivative there
- Can have more than one
- Subdifferential is always a nonempty, compact, convex set
- A convex f is ρ-Lipschitz (on a convex open set) iff all of its subgradients have $\|v\| \leq \rho$

- **Subgradient descent**: instead of a gradient, pick any subgradient
  - The analysis is exactly the same
  - (kind of) what PyTorch/etc do for ReLU functions anyway!
    - For the real details: "On Correctness of Automatic Differentiation for Non-Differentiable Functions"

Projected gradient descent

- What if gradient descent takes us outside of $\mathscr{H}$?

Projected gradient descent

- What if gradient descent takes us outside of $\mathscr{H}$?
    - Remember, $\mathscr{H}$ should be bounded…

Projected gradient descent

- What if gradient descent takes us outside of $\mathscr{H}$?
  - Remember, $\mathscr{H}$ should be bounded…
- **Projected gradient descent**: after each gradient update, project back

Projected gradient descent

- What if gradient descent takes us outside of $\mathscr{H}$?
  - Remember, $\mathscr{H}$ should be bounded…

- **Projected gradient descent**: after each gradient update, project back
  - $w^+ = \text{proj}_{\mathscr{H}}\big(w - \eta \nabla f(w)\big)$ $\quad$ $\text{proj}_{\mathscr{H}}(w) = \text{argmin}_{\hat{w} \in \mathscr{H}} ||\hat{w} - w||$

Projected gradient descent

- What if gradient descent takes us outside of $\mathscr{H}$?
  - Remember, $\mathscr{H}$ should be bounded…

- **Projected gradient descent**: after each gradient update, project back
  - $w^+ = \text{proj}_{\mathscr{H}}\big(w - \eta \nabla f(w)\big) \qquad \text{proj}_{\mathscr{H}}(w) = \text{argmin}_{\hat{w} \in \mathscr{H}} \|\hat{w} - w\|$
  - For $\mathscr{H} = \{w : \|w\| \leq B\}$, $\text{proj}_{\mathscr{H}}(w) = \begin{cases} w & \text{if } \|w\| \leq B \\ \frac{B}{\|w\|} w & \text{otherwise} \end{cases}$

Projected gradient descent

- What if gradient descent takes us outside of $\mathscr{H}$?
  - Remember, $\mathscr{H}$ should be bounded…

- **Projected gradient descent**: after each gradient update, project back
  - $w^+ = \mathrm{proj}_{\mathscr{H}}\big(w - \eta \nabla f(w)\big) \qquad \mathrm{proj}_{\mathscr{H}}(w) = \mathrm{argmin}_{\hat{w} \in \mathscr{H}} \|\hat{w} - w\|$
  - For $\mathscr{H} = \{w : \|w\| \leq B\}$, $\mathrm{proj}_{\mathscr{H}}(w) = \begin{cases} w & \text{if } \|w\| \leq B \\ \frac{B}{\|w\|} w & \text{otherwise} \end{cases}$
  - Analysis is again basically the same

Stochastic gradient descent

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$
    - or, rather, any vector $\hat{v}$ such that $\mathbb{E}[\hat{v} \mid w] \in \partial_w L_{\mathcal{D}}(w)$

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$
    - or, rather, any vector $\hat{v}$ such that $\mathbb{E}[\hat{v} \mid w] \in \partial_w L_{\mathscr{D}}(w)$
  - In general, for an objective $f(w)$ with constraint set $\mathscr{W}$:

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$
    - or, rather, any vector $\hat{v}$ such that $\mathbb{E}[\hat{v} \mid w] \in \partial_w L_{\mathscr{D}}(w)$
  - In general, for an objective $f(w)$ with constraint set $\mathscr{W}$:
    - Start at some $w^{(1)}$, say 0

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$
    - or, rather, any vector $\hat{v}$ such that $\mathbb{E}[\hat{v} \mid w] \in \partial_w L_{\mathscr{D}}(w)$
  - In general, for an objective $f(w)$ with constraint set $\mathscr{W}$:
    - Start at some $w^{(1)}$, say 0
    - Get a random $\hat{g}^{(t)}$ such that $\mathbb{E}\hat{g}^{(t)} \in \partial f(w^{(t)})$

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$
    - or, rather, any vector $\hat{v}$ such that $\mathbb{E}[\hat{v} \mid w] \in \partial_w L_{\mathcal{D}}(w)$
  - In general, for an objective $f(w)$ with constraint set $\mathcal{W}$:
    - Start at some $w^{(1)}$, say 0
    - Get a random $\hat{g}^{(t)}$ such that $\mathbb{E}\hat{g}^{(t)} \in \partial f(w^{(t)})$
    - Set $w^{(t+1)} = \operatorname{proj}_{\mathcal{W}}\left(w^{(t)} - \eta^{(t)} \hat{g}^{(t)}\right)$

Stochastic gradient descent

- **SGD**: why bother computing $L_S(w)$ on the full S every time?
  - Instead, "pure SGD" picks a *fresh* z_i and steps to $w^+ = w - \eta \nabla_w \ell(w, z_i)$
    - or, rather, on a direction in $\partial_w \ell(w, z_i)$
    - or, rather, any vector $\hat{v}$ such that $\mathbb{E}[\hat{v} \mid w] \in \partial_w L_{\mathcal{D}}(w)$
  - In general, for an objective $f(w)$ with constraint set $\mathscr{W}$:
    - Start at some $w^{(1)}$, say 0
    - Get a random $\hat{g}^{(t)}$ such that $\mathbb{E}\hat{g}^{(t)} \in \partial f(w^{(t)})$
    - Set $w^{(t+1)} = \operatorname{proj}_{\mathscr{W}}\left(w^{(t)} - \eta^{(t)} \hat{g}^{(t)}\right)$
    - Return $w^{(T)}$, or $\frac{1}{T} \sum_{t=1}^{T} w^{(t)}$, or whatever

SGD for Lipschitz objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is convex, minimized at $w^* \in \mathcal{H}$, $\sup_{w \in \mathcal{H}} \|w\| \leq B$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/\sqrt{t}$:

$$\mathbb{E}[f(w^{(T)})] - f(w^*) \leq \left(\frac{4B^2}{c} + cG^2\right)\frac{2 + \log T}{\sqrt{T}}$$

SGD for Lipschitz objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is convex, minimized at $w^* \in \mathscr{H}$, $\sup_{w \in \mathscr{H}} \|w\| \leq B$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/\sqrt{t}$:

$$\mathbb{E}[f(w^{(T)})] - f(w^*) \leq \left(\frac{4B^2}{c} + cG^2\right) \frac{2 + \log T}{\sqrt{T}}$$

but…I think the proof might be wrong? or I was too tired last night to understand :(

SGD for Lipschitz objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is convex, minimized at $w^* \in \mathcal{H}$, $\sup_{w \in \mathcal{H}} \|w\| \leq B$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/\sqrt{t}$:

$$\mathbb{E}[f(w^{(T)})] - f(w^*) \leq \left(\frac{4B^2}{c} + cG^2\right) \frac{2 + \log T}{\sqrt{T}}$$

but…I think the proof might be wrong? or I was too tired last night to understand :(

SSBD Theorem 14.8 gives $\mathbb{E}[f(\bar{w})] - f(w^*) \leq \dfrac{B\rho}{\sqrt{T}}$ for $\eta = \dfrac{B}{\rho\sqrt{T}}$, $\bar{w}$ the average

SGD for strongly convex objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is λ-**strongly** convex, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{17cG^2(1 + \log T)}{\lambda T}$$

SGD for strongly convex objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is λ-**strongly** convex, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{17cG^2(1 + \log T)}{\lambda T}$$

- f is **λ-strongly convex** for a parameter $\lambda > 0$ if:
  - $f(\alpha x + (1-\alpha)y) \leq \alpha f(y) + (1-\alpha)f(y) - \frac{1}{2}\lambda\alpha(1-\alpha)\|x - y\|^2$

SGD for strongly convex objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is λ-**strongly** convex, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{17cG^2(1 + \log T)}{\lambda T}$$

- f is **λ-strongly convex** for a parameter $\lambda > 0$ if:
  - $f(\alpha x + (1-\alpha)y) \leq \alpha f(y) + (1-\alpha)f(y) - \frac{1}{2}\lambda\alpha(1-\alpha)\|x-y\|^2$
  - $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \lambda\|x-y\|^2$

SGD for strongly convex objectives

Theorem (<u>Shamir and Zhang, ICML 2013</u>)**:** if f is λ-**strongly** convex, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \le G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \ge 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \le \frac{17cG^2(1 + \log T)}{\lambda T}$$

- f is **λ-strongly convex** for a parameter $\lambda > 0$ if:
    - $f(\alpha x + (1-\alpha)y) \le \alpha f(y) + (1-\alpha)f(y) - \frac{1}{2}\lambda\alpha(1-\alpha)\|x - y\|^2$
    - $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \lambda \|x - y\|^2$
    - $\nabla^2 f \succeq \lambda I$ i.e. $\nabla^2 f - \lambda I \succeq 0$ i.e. all eigenvalues of $\nabla^2 f$ are at least λ

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta c^2 G^2}{\lambda^2 T}$$

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta c^2 G^2}{\lambda^2 T}$$

Can assume $c = 1$ WLOG:

If f is λ-strongly convex, it's also $\frac{\lambda}{c}$-strongly convex;

just use the $c = 1$ theorem with the (weaker) $\frac{\lambda}{c}$ strong convexity param

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = 1/(\lambda t)$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta G^2}{\lambda^2 T}$$

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = 1/(\lambda t)$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta G^2}{\lambda^2 T}$$

Recall key property of β-smoothness: $f(v) \leq f(w) + \langle \nabla f(w), v - w\rangle + \frac{\beta}{2}\|v - w\|^2$

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = 1/(\lambda t)$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta G^2}{\lambda^2 T}$$

Recall key property of β-smoothness: $f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2}\|v - w\|^2$

Plug in w_T, w^*: $f(w_T) - f(w^*) \leq \langle \nabla f(w^*), w_T - w^* \rangle + \frac{\beta}{2}\|w_T - w^*\|^2$

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = 1/(\lambda t)$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta G^2}{\lambda^2 T}$$

Recall key property of β-smoothness: $f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2}\|v - w\|^2$

Plug in w_T, w^*: $f(w_T) - f(w^*) \leq \langle \nabla f(w^*), w_T - w^* \rangle + \frac{\beta}{2}\|w_T - w^*\|^2$

$$= \frac{\beta}{2}\|w_T - w^*\|^2$$

SGD for strongly convex objectives

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = 1/(\lambda t)$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta G^2}{\lambda^2 T}$$

Recall key property of β-smoothness: $f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2}\|v - w\|^2$

Plug in w_T, w^*:
$$\begin{aligned} f(w_T) - f(w^*) &\leq \langle \nabla f(w^*), w_T - w^* \rangle + \frac{\beta}{2}\|w_T - w^*\|^2 \\ &= \frac{\beta}{2}\|w_T - w^*\|^2 \end{aligned}$$

Lemma: if f is λ-strongly convex, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = 1/(\lambda t)$, then $\mathbb{E}\left[\|w_T - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 T}$.

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right]$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] = \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right]$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\ &\le \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right]\end{aligned}$$

since $\mathscr{W}$ is convex

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$
\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\
&\leq \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] \qquad \text{since } \mathscr{W} \text{ is convex} \\
&= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right]
\end{aligned}
$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$
\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\text{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\
&\leq \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] \qquad \text{since } \mathscr{W} \text{ is convex} \\
&= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right]
\end{aligned}
$$

$$
\mathbb{E}[\langle \hat{g}_t, w_t - w^* \rangle] = \mathbb{E}_{w_t}\left[\mathbb{E}_{\hat{g}_t}\left[\langle \hat{g}_t, w_t - w^* \rangle \mid w_t\right]\right]
$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\ &\leq \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] \qquad \text{since } \mathscr{W} \text{ is convex} \\ &= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\langle \hat{g}_t, w_t - w^* \rangle] &= \mathbb{E}_{w_t}\left[\mathbb{E}_{\hat{g}_t}\left[\langle \hat{g}_t, w_t - w^* \rangle \mid w_t\right]\right] \\ &= \mathbb{E}_{w_t}\left[\langle g_t, w_t - w^* \rangle\right] \qquad \text{for some } g_t \in \partial f(w_t)\end{aligned}$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$
\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\
&\leq \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] \qquad \text{since } \mathscr{W} \text{ is convex} \\
&= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\langle \hat{g}_t, w_t - w^* \rangle] &= \mathbb{E}_{w_t}\left[\mathbb{E}_{\hat{g}_t}\left[\langle \hat{g}_t, w_t - w^* \rangle \mid w_t\right]\right] \\
&= \mathbb{E}_{w_t}\left[\langle g_t, w_t - w^* \rangle\right] \qquad \text{for some } g_t \in \partial f(w_t) \\
&= \mathbb{E}_{w_t}\left[\langle g_t - \nabla f(w^*), w_t - w^* \rangle\right]
\end{aligned}
$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$
\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\
&\leq \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] && \text{since } \mathscr{W} \text{ is convex} \\
&= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\langle \hat{g}_t, w_t - w^* \rangle] &= \mathbb{E}_{w_t}\left[\mathbb{E}_{\hat{g}_t}\left[\langle \hat{g}_t, w_t - w^* \rangle \mid w_t\right]\right] \\
&= \mathbb{E}_{w_t}\left[\langle g_t, w_t - w^* \rangle\right] && \text{for some } g_t \in \partial f(w_t) \\
&= \mathbb{E}_{w_t}\left[\langle g_t - \nabla f(w^*), w_t - w^* \rangle\right] \\
&\geq \lambda \|w_t - w^*\|^2 && \text{first-order strong convexity def}
\end{aligned}
$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\
&\le \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] && \text{since } \mathscr{W} \text{ is convex} \\
&= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right] \\
&\le (1 - 2\eta_t \lambda)\mathbb{E}\left[\|w_t - w^*\|^2\right] + \eta_t^2 G^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\langle \hat{g}_t, w_t - w^* \rangle] &= \mathbb{E}_{w_t}\left[\mathbb{E}_{\hat{g}_t}\left[\langle \hat{g}_t, w_t - w^* \rangle \mid w_t\right]\right] \\
&= \mathbb{E}_{w_t}\left[\langle g_t, w_t - w^* \rangle\right] && \text{for some } g_t \in \partial f(w_t) \\
&= \mathbb{E}_{w_t}\left[\langle g_t - \nabla f(w^*), w_t - w^* \rangle\right] \\
&\ge \lambda \|w_t - w^*\|^2 && \text{first-order strong convexity def}
\end{aligned}$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathscr{W}}(w_t - \eta_t \hat{g}_t) - w^*\|^2\right] \\
&\le \mathbb{E}\left[\|w_t - \eta_t \hat{g}_t - w^*\|^2\right] \qquad \text{since } \mathscr{W} \text{ is convex} \\
&= \mathbb{E}\left[\|w_t - w^*\|^2\right] - 2\eta_t \mathbb{E}\left[\langle \hat{g}_t, w_t - w^* \rangle\right] + \eta_t^2 \mathbb{E}\left[\|\hat{g}_t\|^2\right] \\
&\le (1 - 2\eta_t \lambda)\mathbb{E}\left[\|w_t - w^*\|^2\right] + \eta_t^2 G^2 \\
&= \left(1 - \frac{2}{t}\right)\mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\langle \hat{g}_t, w_t - w^* \rangle] &= \mathbb{E}_{w_t}\left[\mathbb{E}_{\hat{g}_t}\left[\langle \hat{g}_t, w_t - w^* \rangle \mid w_t\right]\right] \\
&= \mathbb{E}_{w_t}\left[\langle g_t, w_t - w^* \rangle\right] \qquad \text{for some } g_t \in \partial f(w_t) \\
&= \mathbb{E}_{w_t}\left[\langle g_t - \nabla f(w^*), w_t - w^* \rangle\right] \\
&\ge \lambda \|w_t - w^*\|^2 \qquad \text{first-order strong convexity def}
\end{aligned}$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \le \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{G^2}{\lambda^2}$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{G^2}{\lambda^2}$$

implies $\mathbb{E}\left[\|w_1 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 1}$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \le \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \frac{G^2}{\lambda^2}$$

implies $\mathbb{E}\left[\|w_1 - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 \cdot 1}$ $\quad$ $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 \cdot 2}$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{G^2}{\lambda^2}$$

implies $\mathbb{E}\left[\|w_1 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 1}$ $\quad$ $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 2}$

induction for $t \geq 3$: have

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{G^2}{\lambda^2}$$

implies $\mathbb{E}\left[\|w_1 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 1}$ $\quad$ $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 2}$

induction for $t \geq 3$: have

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \frac{4G^2}{\lambda^2 t} + \frac{G^2}{\lambda^2 t^2}$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \le G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \le \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \frac{G^2}{\lambda^2}$$

implies $\mathbb{E}\left[\|w_1 - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 \cdot 1}$ $\quad \mathbb{E}\left[\|w_2 - w^*\|^2\right] \le \frac{4G^2}{\lambda^2 \cdot 2}$

induction for $t \ge 3$: have

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \le \left(1 - \frac{2}{t}\right) \frac{4G^2}{\lambda^2 t} + \frac{G^2}{\lambda^2 t^2} = \frac{G^2}{\lambda^2}\left[\frac{4}{t} - \frac{8}{t^2} + \frac{1}{t^2}\right]$$

Assumptions: f λ-strongly convex, $\mathbb{E}\left[\|g\|^2\right] \leq G^2$, $w^{(0)} = 0$, $\eta^{(t)} = 1/(\lambda t)$

WTS $\mathbb{E}\left[\|w_t - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 t}$

have $\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{G^2}{\lambda^2 t^2}$

plugging in $t = 1$, $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \left(1 - \frac{2}{1}\right) \mathbb{E}\left[\|w_1 - w^*\|^2\right] + \frac{G^2}{\lambda^2 \cdot 1^2}$

$$\mathbb{E}\left[\|w_1 - w^*\|^2\right] + \mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{G^2}{\lambda^2}$$

implies $\mathbb{E}\left[\|w_1 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 1}$ $\quad$ $\mathbb{E}\left[\|w_2 - w^*\|^2\right] \leq \frac{4G^2}{\lambda^2 \cdot 2}$

induction for $t \geq 3$: have

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - \frac{2}{t}\right) \frac{4G^2}{\lambda^2 t} + \frac{G^2}{\lambda^2 t^2} = \frac{G^2}{\lambda^2}\left[\frac{4}{t} - \frac{8}{t^2} + \frac{1}{t^2}\right] \leq \frac{G^2}{\lambda^2}\left[\frac{4}{t+1}\right]$$

SGD for strongly convex objectives

Theorem (Shamir and Zhang, ICML 2013)**:** if f is λ-**strongly** convex, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{17cG^2(1 + \log T)}{\lambda T}$$

Proof uses that lemma (which doesn't need β-smoothness) as a key step, but does some more tricks – read it if you're interested!

Implications for learning

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta c^2 G^2}{\lambda^2 T}$$

Implications for learning

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta c^2 G^2}{\lambda^2 T}$$

- So, if:
  - $L_{\mathscr{D}}$ is λ-strongly convex
  - $L_{\mathscr{D}}$ is β-smooth (e.g. implied if $\ell(\cdot, z)$ is β-smooth)
  - $\mathbb{E}\left[\|\hat{g}_t\|^2\right] \leq G^2$ (e.g. implied if $\ell(\cdot, z)$ is G-Lipschitz)
  - minimizer is inside $\mathscr{H}$

Implications for learning

Theorem: if f is λ-strongly convex and β-smooth, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \frac{2\beta c^2 G^2}{\lambda^2 T}$$

- So, if:
  - $L_{\mathscr{D}}$ is λ-strongly convex
  - $L_{\mathscr{D}}$ is β-smooth (e.g. implied if $\ell(\cdot, z)$ is β-smooth)
  - $\mathbb{E}\left[\|\hat{g}_t\|^2\right] \leq G^2$ (e.g. implied if $\ell(\cdot, z)$ is G-Lipschitz)
  - minimizer is inside $\mathscr{H}$
- then we have a bound on expected excess error for SGD
  - Needs $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ samples, since this analysis is for one-pass only

Implications for learning

SSBD Theorem 14.11: if f is λ-strongly convex, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{cG^2}{2\lambda T}(1 + \log(T))$$

Implications for learning

SSBD Theorem 14.11: if f is λ-strongly convex, minimized at $w^* \in \mathcal{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{cG^2}{2\lambda T}(1 + \log(T))$$

- So, if:
  - $L_{\mathcal{D}}$ is λ-strongly convex
  - $\mathbb{E}\left[\|\hat{g}_t\|^2\right] \leq G^2$ (e.g. implied if $\ell(\cdot, z)$ is G-Lipschitz)
  - minimizer is inside $\mathcal{H}$

Implications for learning

SSBD Theorem 14.11: if f is λ-strongly convex, minimized at $w^* \in \mathscr{H}$, $\mathbb{E}\left[\|\hat{g}^{(t)}\|^2\right] \leq G^2$ for all t, and $\eta^{(t)} = c/(\lambda t)$ for $c \geq 1$:

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{cG^2}{2\lambda T}(1 + \log(T))$$

- So, if:
  - $L_{\mathscr{D}}$ is λ-strongly convex
  - $\mathbb{E}\left[\|\hat{g}_t\|^2\right] \leq G^2$ (e.g. implied if $\ell(\cdot, z)$ is G-Lipschitz)
  - minimizer is inside $\mathscr{H}$
- then we have a bound on expected excess error for average iterate of SGD
  - Needs $\mathscr{O}\left(\frac{1}{\varepsilon}\right)$ samples, since this analysis is for one-pass only

Implications for learning

SSBD Theorem 14.8: if f is convex, $\mathscr{H} = \{w : \|w\| \leq B\}$, $w^* \in \operatorname{argmin}_{w \in \mathscr{H}} f(w)$, $\Pr(\|\hat{g}_t\| \leq \rho) = 1$ for all t, and $\eta = B/(\rho\sqrt{T})$, then

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Implications for learning

SSBD Theorem 14.8: if f is convex, $\mathscr{H} = \{w : \|w\| \leq B\}$, $w^* \in \operatorname{argmin}_{w \in \mathscr{H}} f(w)$, $\Pr(\|\hat{g}_t\| \leq \rho) = 1$ for all t, and $\eta = B/(\rho\sqrt{T})$, then

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

- So, if:
  - $L_{\mathscr{D}}$ is convex (e.g. implied if $\ell(\cdot, z)$ is convex)
  - $\|\hat{g}_t\| \leq \rho$ a.s. (e.g. implied if $\ell(\cdot, z)$ is G-Lipschitz)

Implications for learning

SSBD Theorem 14.8: if f is convex, $\mathcal{H} = \{w : \|w\| \leq B\}$, $w^* \in \operatorname{argmin}_{w \in \mathcal{H}} f(w)$, $\Pr(\|\hat{g}_t\| \leq \rho) = 1$ for all t, and $\eta = B/(\rho\sqrt{T})$, then

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

- So, if:
  - $L_{\mathcal{D}}$ is convex (e.g. implied if $\ell(\cdot, z)$ is convex)
  - $\|\hat{g}_t\| \leq \rho$ a.s. (e.g. implied if $\ell(\cdot, z)$ is G-Lipschitz)
- then we have a bound on expected excess error for SGD
  - Needs $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ samples, since this analysis is for one-pass only
  - Covers the Convex-Lipschitz-Bounded case

Implications for learning

SSBD Theorem 14.13: if $\ell(\cdot, z)$ is convex, β-smooth, and nonnegative, $\mathscr{H} = \{w : \|w\| \leq B\}$, and η is constant, then for any w^*

$$\mathbb{E}[L_{\mathscr{D}}(\bar{w})] \leq \frac{1}{1-\eta\beta}\left(L_{\mathscr{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T}\right).$$

Implications for learning

SSBD Theorem 14.13: if $\ell(\cdot, z)$ is convex, β-smooth, and nonnegative, $\mathcal{H} = \{w : \|w\| \leq B\}$, and η is constant, then for any w^*

$$\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \frac{1}{1-\eta\beta}\left(L_{\mathcal{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T}\right).$$

- So, if we take $\eta = 1/(\beta(1+3/\varepsilon))$, $T \geq 12B^2\beta^2/\varepsilon^2$, and assume $\ell(0,z) \leq 1$, $\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \min_{w \in \mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon$.

Implications for learning

SSBD Theorem 14.13: if $\ell(\cdot, z)$ is convex, β-smooth, and nonnegative, $\mathcal{H} = \{w : \|w\| \leq B\}$, and η is constant, then for any w^*

$$\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \frac{1}{1-\eta\beta}\left(L_{\mathcal{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T}\right).$$

- So, if we take $\eta = 1/(\beta(1+3/\varepsilon))$, $T \geq 12B^2\beta^2/\varepsilon^2$, and assume $\ell(0,z) \leq 1$, $\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \leq \min_{w \in \mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon$.
  - Covers the Convex-Smooth-Bounded case

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
    - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…
  - But it's *necessary* in this framework

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…
  - But it's *necessary* in this framework
    - ERM with gradient descent does *not* always work

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…
  - But it's *necessary* in this framework
    - ERM with gradient descent does *not* always work
- "Early stopping" with one-pass SGD is a form of **(implicit) regularization**

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…
  - But it's *necessary* in this framework
    - ERM with gradient descent does *not* always work
- "Early stopping" with one-pass SGD is a form of **(implicit) regularization**
- After the break, *explicit* regularization:

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…
  - But it's *necessary* in this framework
    - ERM with gradient descent does *not* always work
- "Early stopping" with one-pass SGD is a form of **(implicit) regularization**
- After the break, *explicit* regularization:
  - makes things strongly convex

Summary

- One-pass SGD can always learn convex, Lipschitz/smooth, bounded problems
- Rate is better with strong convexity
  - $\mathcal{O}(1/n)$ excess error, vs $\mathcal{O}(1/\sqrt{n})$ without
  - For gradient descent, the gap is enormous: $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps vs $\mathcal{O}\left(\log\frac{1}{\varepsilon}\right)$
- We didn't analyze multi-pass SGD
  - Only looking at each data point once might be wasteful…
  - But it's *necessary* in this framework
    - ERM with gradient descent does *not* always work
- "Early stopping" with one-pass SGD is a form of **(implicit) regularization**
- After the break, *explicit* regularization:
  - makes things strongly convex
  - lets us learn even if we fully optimize on S