

SESSION 16: STATISTICAL MACHINE LEARNING (VI)



Gang Li

Deakin University, Australia

2021-10-09

Convex Optimisation	3
Properties of Convexity	7
Convex Optimization Algorithms	12
Convex Optimization	13
Gradient Descent: first-order method.	14
Analysing Gradient Descent for Lipschitz functions	15
Stochastic Gradient Descent (SGD): first-order method	16
Analysing Stochastic Gradient Descent for Lipschitz functions	17
Newton’s method: second-order method.	18
Convex Learning Problems	19
Convex Learning Problems	20
Convex-Lipschitz-bounded learning problem.	21
Convex-Smooth-bounded learning problem	22
Surrogate Loss Functions	23
Surrogate Loss Functions	24
The 0 – 1 Loss Function and Hinge Loss	25
Error Decomposition Revisited	26
Quiz	27

Table of Content

Convex Optimisation

Properties of Convexity

Convex Optimization Algorithms

Convex Optimization

Gradient Descent: first-order method

Analysing Gradient Descent for Lipschitz functions

Stochastic Gradient Descent (SGD): first-order method

Analysing Stochastic Gradient Descent for Lipschitz functions

Newton’s method: second-order method

Convex Learning Problems

Convex Learning Problems

Convex-Lipschitz-bounded learning problem

Convex-Smooth-bounded learning problem

Surrogate Loss Functions

Surrogate Loss Functions

The 0 – 1 Loss Function and Hinge Loss

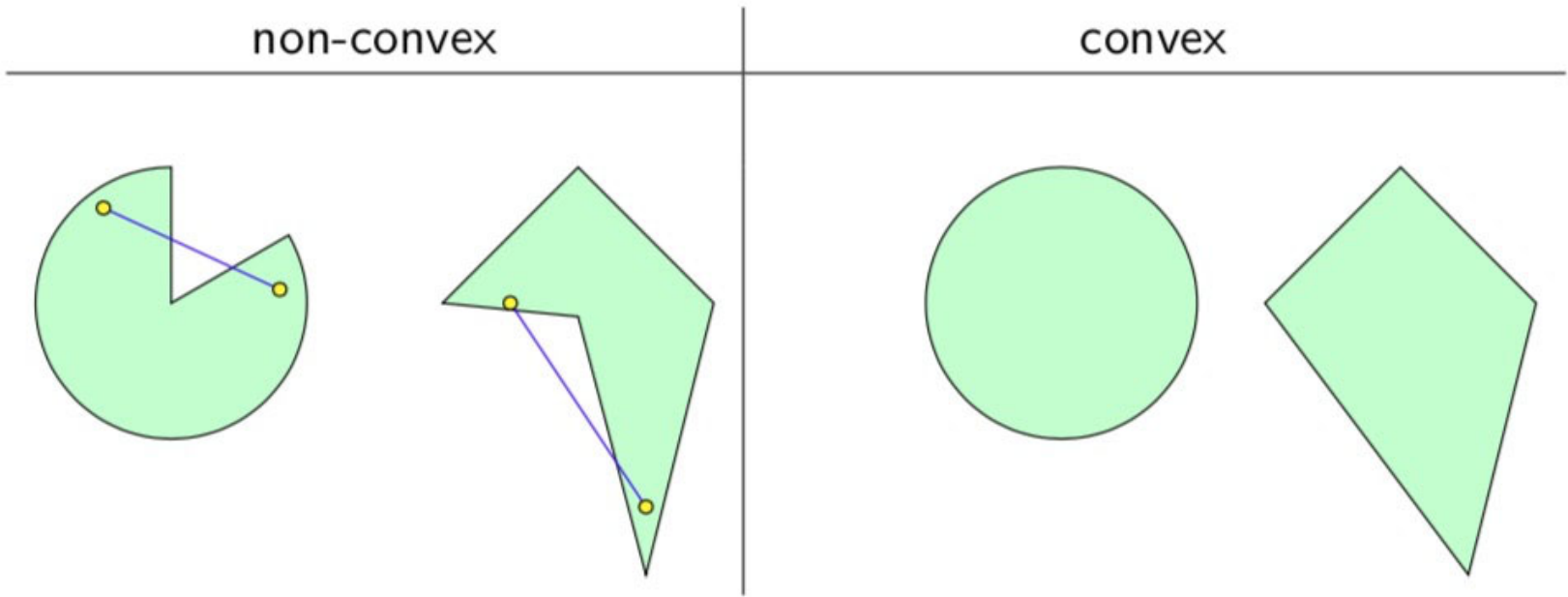
Error Decomposition Revisited

Quiz

Convex Optimisation

Convexity

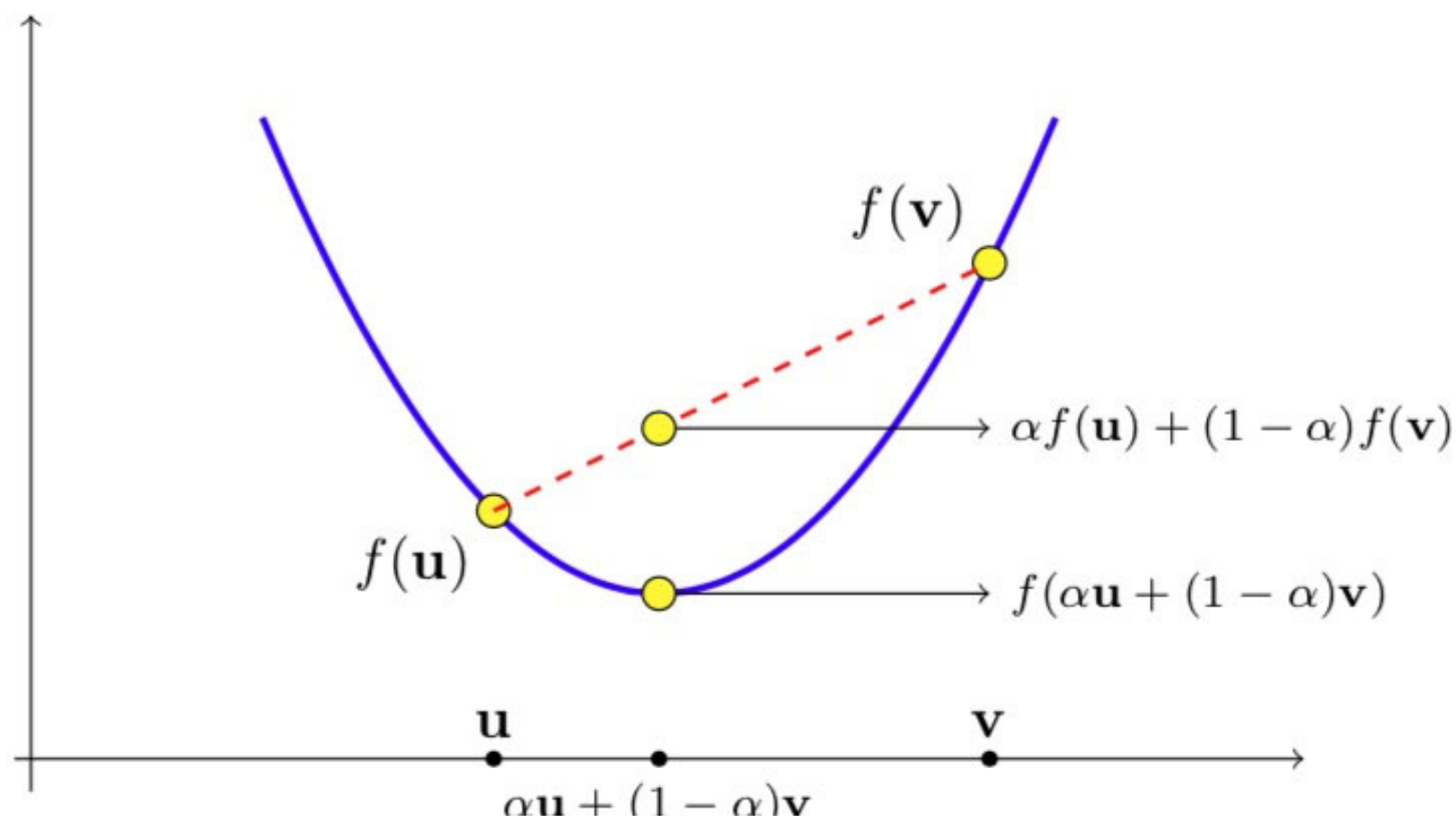
👉 A set C in a vector space is **convex** if for any two vectors $\mathbf{u}, \mathbf{v} \in C$, the line segment between \mathbf{u} and \mathbf{v} is contained in C . Namely, for any $\alpha \in [0, 1]$ we have the **convex combination** $\alpha \mathbf{u} + (1 - \alpha) \mathbf{v}$ is in C



Convexity

Let set C be a convex set. A function $f : C \mapsto \mathbb{R}$ is **convex** if for any two vectors $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

$f(\alpha \mathbf{u} + (1 - \alpha) \mathbf{v}) \leq \alpha f(\mathbf{u}) + (1 - \alpha) f(\mathbf{v})$

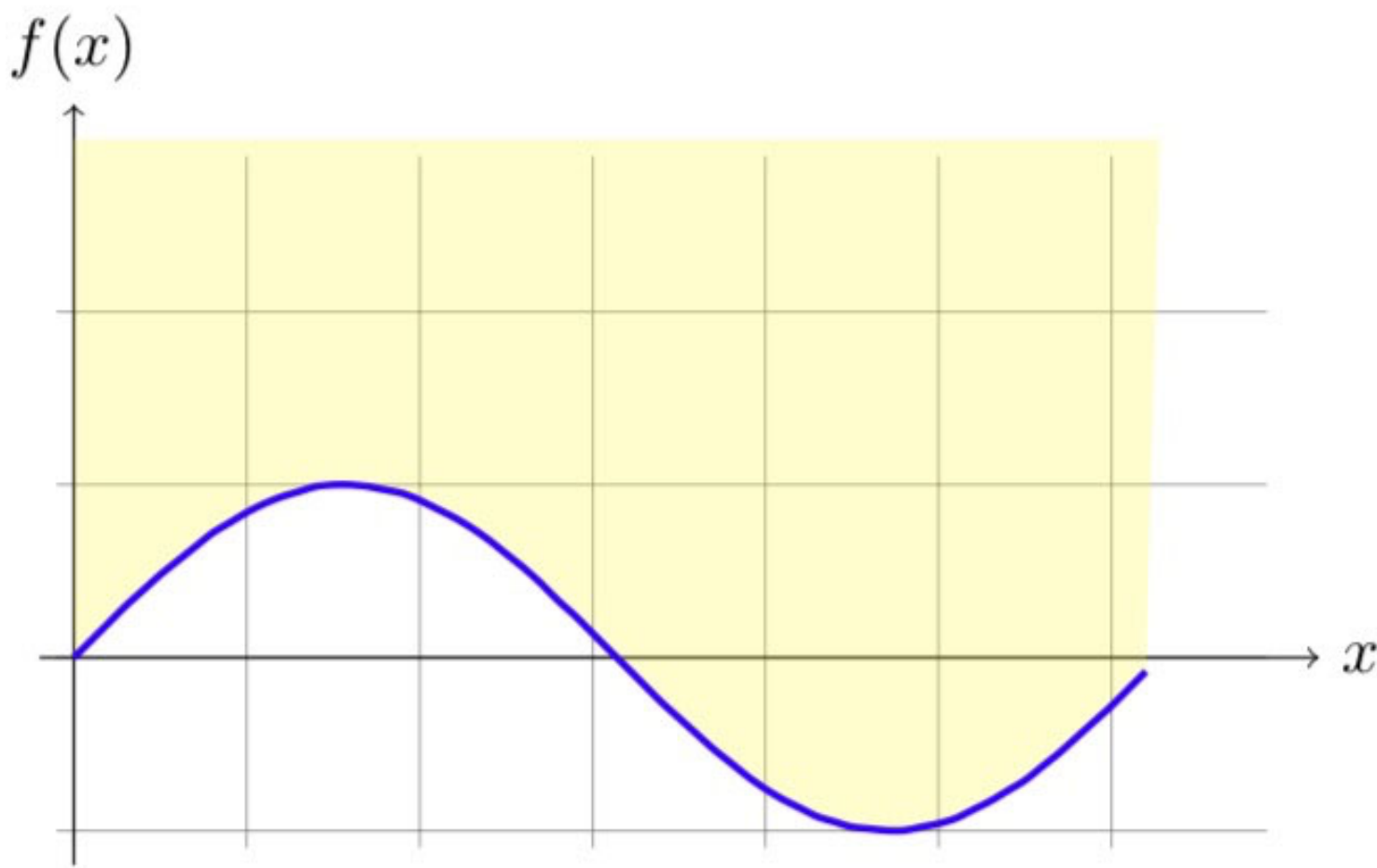


(None)-4f977b3 (2021-10-09) – 5 / 30

Convexity

A function f is **convex** if and only if its **epigraph** is a convex set:

$epigraph(f) = \{(\mathbf{x}, \beta) : f(\mathbf{x}) \leq \beta\}$



(None)-4f977b3 (2021-10-09) – 6 / 30

Properties of Convexity (I)

👉 If a function f is **convex**, then every **local minimum** of f is also a **global minimum**.

Claims.

- Let $B(u, r) = \{v : \|v - u\| \leq r\}$
- $f(u)$ is a local minimum of f at u if $\exists r > 0$, s.t. $\forall v \in B(u, r)$ we have $f(v) \geq f(u)$
- It follows that for any v (not necessarily in B), there is a small enough $\alpha > 0$ such that $u + \alpha(v - u) \in B(u, r)$ and therefore
$$f(\mathbf{u}) \leq f(\mathbf{u} + \alpha(\mathbf{u} - \mathbf{v}))$$
- If f is convex, we also have that
$$f(\mathbf{u} + \alpha(\mathbf{u} - \mathbf{v})) = f(\alpha\mathbf{v} + (1 - \alpha)\mathbf{u}) \leq (1 - \alpha)f(\mathbf{u}) + \alpha f(\mathbf{v})$$
- Combining and rearranging terms, we obtain that $f(\mathbf{u}) \leq f(\mathbf{v})$
- This holds for every \mathbf{v} , hence $f(\mathbf{u})$ is also a global minimum of f

□

(None)-4f977b3 (2021-10-09) – 7 / 30

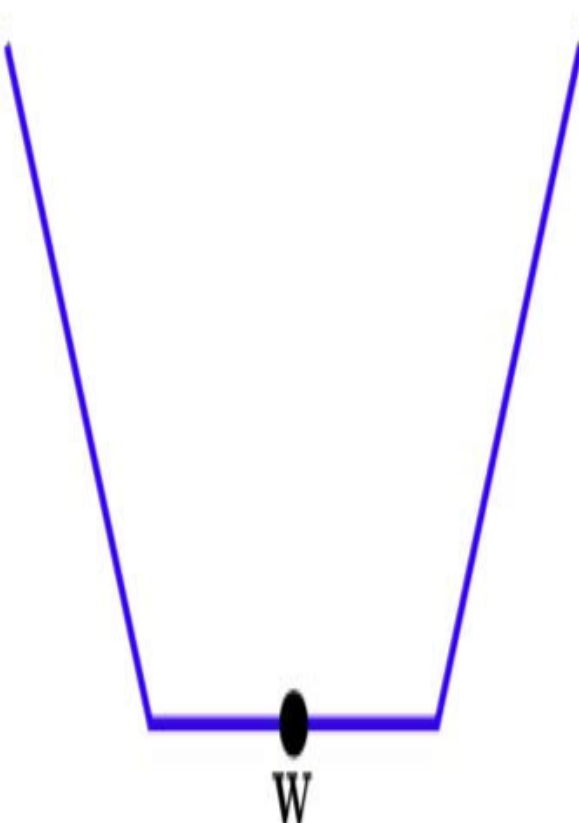
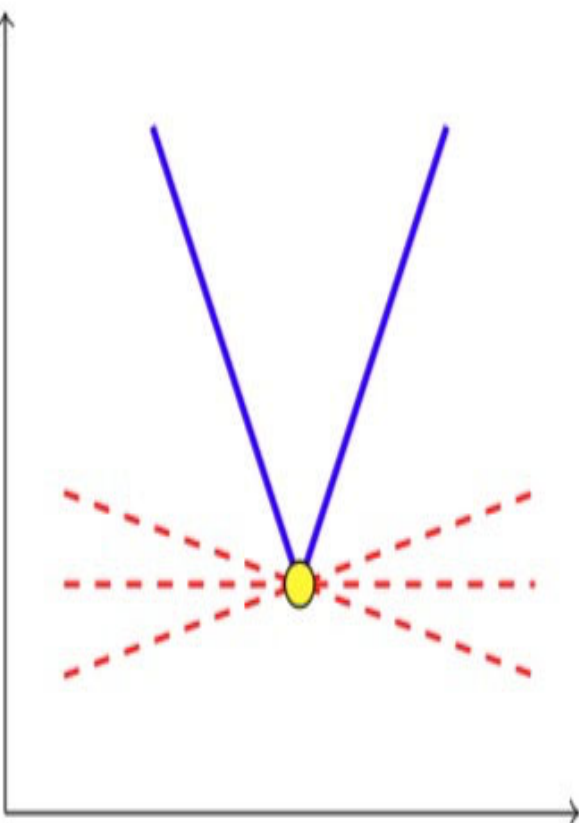
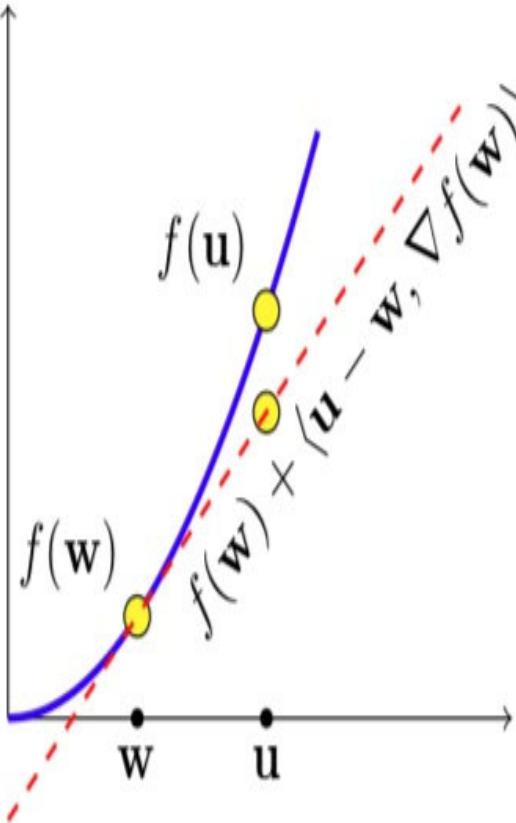
Properties of Convexity (II)

If a function f is **convex** and differentiable, then

$\forall \mathbf{u}, f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$

👉

- $\nabla f(\mathbf{w}) = (\frac{\partial f(\mathbf{w})}{\partial \omega_1}, \dots, \frac{\partial f(\mathbf{w})}{\partial \omega_d})$ is the **gradient** of f at \mathbf{w}
- A vector \mathbf{v} is a **sub-gradient** of f at \mathbf{w} if $\forall \mathbf{u}, f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle$
- The set of sub-gradients of f at \mathbf{w} is called the **differential set**, $\partial f(\mathbf{w})$.
 - ◆ f is convex if and only if for every \mathbf{w} , $\partial f(\mathbf{w}) \neq \emptyset$
 - ◆ f is “**locally flat**” around \mathbf{w} ($0 \in \partial f(\mathbf{w})$) iff \mathbf{w} is a global minimiser.



(None)-4f977b3 (2021-10-09) – 8 / 30

Properties of Convexity (III)

Let f be a scalar twice differential function, and let f' , f'' be its first and second derivatives, respectively, then the following are equivalent

- f is convex
- f' is monotonically non-decreasing
- f'' is non-negative

Claims.

- The composition of a convex scalar function with a linear function yields a convex vector-valued function:
 - ◆ Assume that $f : \mathcal{R}^d \mapsto \mathcal{R}$ can be written as $f(\mathbf{w}) = g(\langle \mathbf{w}, x \rangle + y)$, for some $\mathbf{x} \in \mathcal{R}^d$, $y \in \mathcal{R}$ and $g : \mathcal{R} \mapsto \mathcal{R}$. Then the convexity of g implies the convexity of f .
- The maximum of convex functions is convex and that a weighted sum of convex functions, with non-negative weights, is also convex.
 - ◆ For $i = 1, \dots, r$, let $f_i : \mathcal{R}^d \mapsto \mathcal{R}$ be a convex function. The following functions from \mathcal{R}^d to \mathcal{R} are also convex:
 - $g(x) = \max_{i \in [r]} f_i(x)$
 - $g(x) = \sum_{i=1}^r \omega_i f_i(x)$, where $\forall i, \omega_i \geq 0$.

□

Lipschitzness

A function $f : C \mapsto \mathcal{R}$ is **ρ -Lipschitz** if for every $\omega_1, \omega_2 \in \mathcal{R}$, we have that

$$|f(\omega_1) - f(\omega_2)| \leq \rho \|\omega_1 - \omega_2\|$$

Examples.

- The linear function $f(\omega) = \langle v, \omega \rangle + b$, where $v \in \mathcal{R}^d$ is $\|v\|$ -Lipschitz, from the Cauchy-Schwartz inequality:
$$|f(\omega_1) - f(\omega_2)| = |\langle v, \omega_1 - \omega_2 \rangle| \leq \|v\| \times \|\omega_1 - \omega_2\|$$

□


Claims.

- A Lipschitzness function cannot change too fast: if a Lipschitzness function f is differentiable, then its derivative is everywhere bounded by ρ .
- Let $f(x) = g_1(g_2(x))$, where g_1 is ρ_1 -Lipschitz and g_2 is ρ_2 -Lipschitz. Then f is $\rho_1\rho_2$ -Lipschitz.

□

Smoothness

A differentiable function $f : C \mapsto \mathcal{R}$ is **β -smooth** if its gradient is β -Lipschitz; namely, for all ω_1, ω_2 we have


$$|\nabla f(\omega_1) - \nabla f(\omega_2)| \leq \beta \|\omega_1 - \omega_2\|$$

Claims.

- Smoothness implies that for all v and ω ,

$$f(v) \leq f(\omega) + \langle \nabla f(\omega), v - \omega \rangle + \frac{\beta}{2} \|v - \omega\|^2$$

Recall that convexity implies that $f(v) \geq f(\omega) + \langle \nabla f(\omega), v - \omega \rangle$. Therefore, when a function is both **convex** and **smooth**, we have both **upper** and **lower** bounds on the difference between the function and its first order approximation.

- Setting $v = \omega - \frac{1}{\beta} \nabla f(\omega)$, we have $\frac{1}{2\beta} \|\nabla f(\omega)\|^2 \leq f(\omega) - f(v)$. When $f(v) \geq 0$, we have $\|\nabla f(\omega)\|^2 \leq 2\beta f(\omega)$, a function satisfying this property is called **self-bounded function**.

□

Convex Optimization Algorithms

Convex Optimization

Approximately solve the problem of


$$\operatorname{argmin}_{\omega \in C} f(\omega)$$

where C is a convex set and f is a convex function

Special Cases.

Feasibility f is a constant function

Unconstrained minimization $C = \mathcal{R}^d$

- They can reduce one to another
 - Adding the function $IC(\omega)$ to the objective eliminates the constraint
 - Adding the constraint $f(\omega) \leq f^* + \epsilon$ eliminates the objective

□

Gradient Descent: first-order method

To minimize a differentiable convex function $f(\omega)$, assume the gradient of $f : \mathcal{R}^d \mapsto \mathcal{R}$ at ω , denoted as $\nabla f(\omega) = (\frac{\partial f(\omega)}{\partial \omega_1}, \dots, \frac{\partial f(\omega)}{\partial \omega_d})$ is the **gradient** of f at ω .

■ Start with initial $\omega^{(1)}$ (usually, the zero vector)

■ At iteration t , update $\omega^{(t+1)} = \omega^{(t)} - \eta \nabla f(\omega^{(t)})$ where η is the **learning rate**.

■ **Sub-gradient Descent**: for non differentiable function f , we can replace gradients with sub-gradients: update $\omega^{(t+1)} = \omega^{(t)} - \eta v_t$ where $v_t \in \partial f(\omega^{(t)})$.

Intuition.

- By Taylor’s approximation, if close to $\omega^{(t)}$, we have $f(u) \cong f(\omega^{(t)}) + \langle \nabla f(\omega^{(t)}), u - \omega^{(t)} \rangle$
- We can minimize the appropriation of $f(\omega)$, but also wish the approximation to be accurate. Therefore, we would like to minimize jointly the distance between ω and $\omega^{(t)}$, and the approximation of f around $\omega^{(t)}$. If η controls the trade-off between those two terms, we obtain the update rule:

$$\omega^{(t+1)} = \operatorname{argmin}_{\omega} \frac{1}{2} \|\omega - \omega^{(t)}\|^2 + \eta (f(\omega^{(t)}) + \langle \nabla f(\omega^{(t)}), \omega - \omega^{(t)} \rangle)$$

□

Analysing Gradient Descent for Lipschitz functions

To minimize a convex function $f(\omega)$, let ω^* be any vector, which could be the minimizer of $f(\omega)$, and B be an upper bound on $\|\omega^*\|$, namely $\|\omega^*\| \leq B$. We can obtain an upper bound on the suboptimality w.r.t. ω^* : $f(\bar{\omega}) - f(\omega^*)$, where $\bar{\omega} = \frac{1}{T} \sum_{t=1}^T \omega^{(t)}$.

■

$$f(\bar{\omega}) - f(\omega^*) \leq \frac{1}{T} \sum_{t=1}^T \langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle \leq \frac{B^2}{2\eta T} + \frac{\eta}{2T} \sum_{t=1}^T \|\mathbf{v}_t\|^2$$

Intuition.

- Using Jensen’s inequality, we have

$$f(\bar{\omega}) - f(\omega^*) = f\left(\frac{1}{T} \sum_{t=1}^T \omega^{(t)}\right) - f(\omega^*) \leq \frac{1}{T} \sum_{t=1}^T f(\omega^{(t)}) - f(\omega^*) = \frac{1}{T} \sum_{t=1}^T (f(\omega^{(t)}) - f(\omega^*))$$

- Followed by the convexity of f , we have $f(\omega^{(t)}) - f(\omega^*) \leq \langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle$

□

Intuition.

- Using algebraic manipulations, we have

$$\langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle = \frac{1}{\eta} \langle \eta \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle = \frac{1}{2\eta} (-\|\omega^{(t+1)} - \omega^*\|^2 + \|\omega^{(t)} - \omega^*\|^2) + \frac{\eta}{2} \|\mathbf{v}_t\|^2$$

- Summing them together as below: when $\omega^{(1)} = 0$, the previous statement is proved.

$$\sum_{t=1}^T \langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle = \frac{1}{2\eta} (\|\omega^{(1)} - \omega^*\|^2 - \|\omega^{(T+1)} - \omega^*\|^2) + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2 \leq \frac{1}{2\eta} (\|\omega^{(1)} - \omega^*\|^2) + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2$$

□

Intuition.

- Since f is convex and ρ -Lipschitz, $\|\mathbf{v}_t\| \leq \rho$ for every t . Therefore

$$\frac{1}{T} \sum_{t=1}^T (f(\omega^{(t)}) - f(\omega^*)) \leq \frac{\|\omega^*\|^2}{2\eta T} + \frac{\eta}{2} \rho^2 \leq \frac{B^2}{2\eta T} + \frac{\eta}{2} \rho^2$$

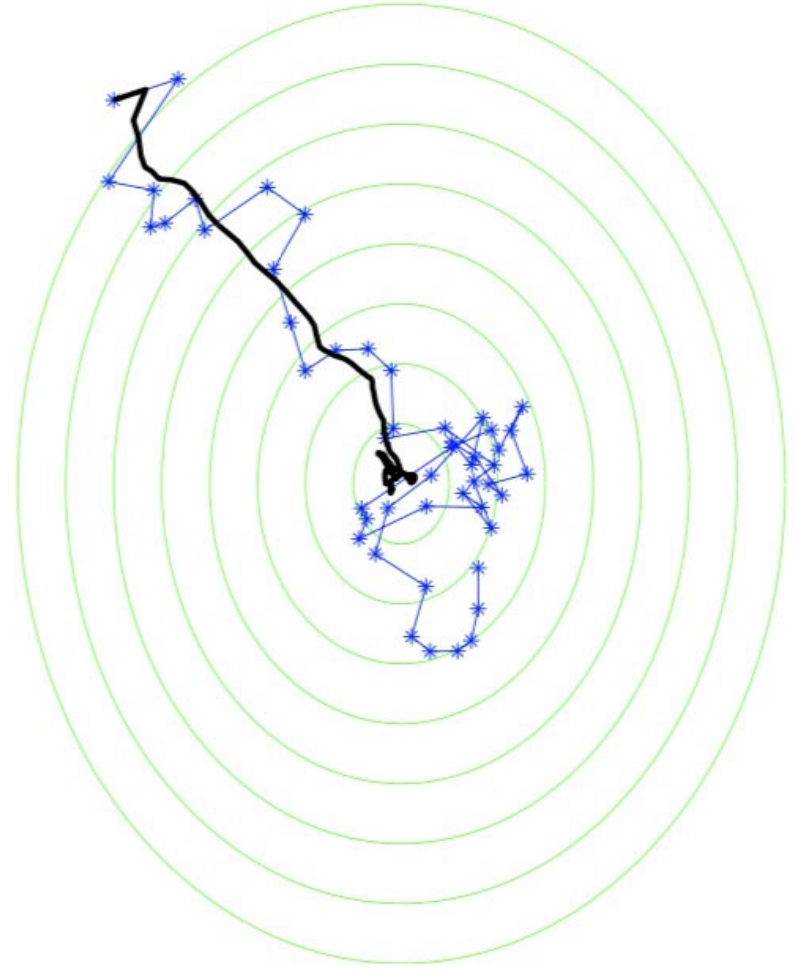
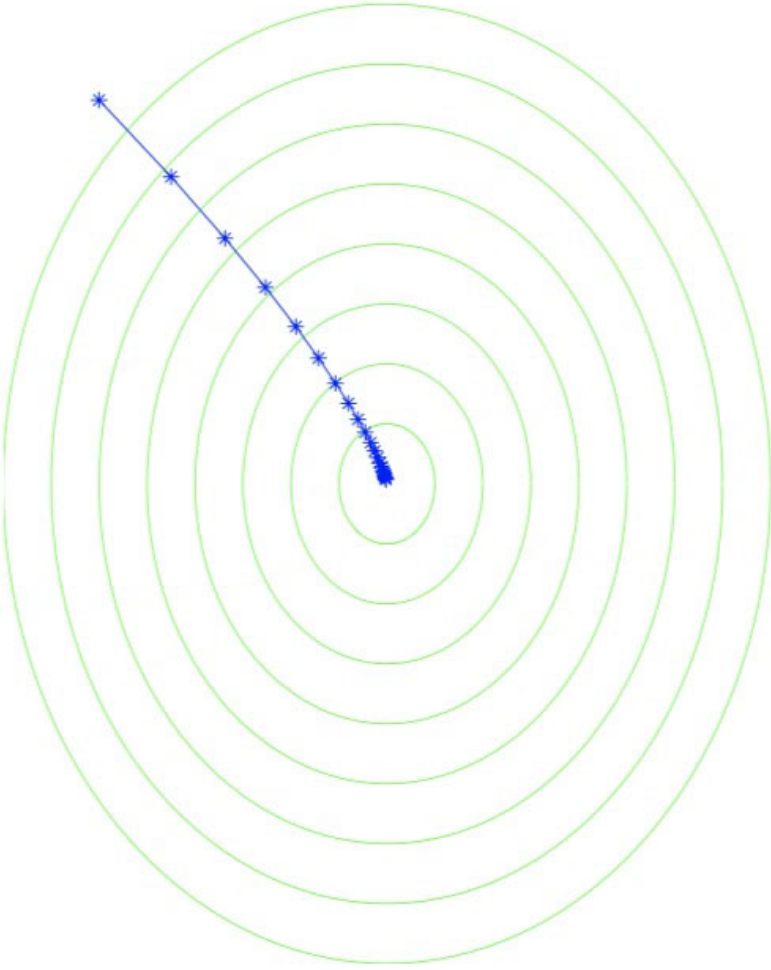
- For every ω^* , if $T > \frac{B^2 \rho^2}{\epsilon^2}$, and $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the right side is at most ϵ . Hence $f(\bar{\omega}) - f(\omega^*) \leq \epsilon$.
- This indicates that the gradient descent method needs $\frac{B^2 \rho^2}{\epsilon^2}$ steps to converge.

□

Stochastic Gradient Descent (SGD): first-order method

To minimize a differentiable convex-Lipschitz-bounded function $f(\omega)$, the goal is to probably approximately solve $\min_{\omega} L_{\mathcal{D}}(\omega)$ where $L_{\mathcal{D}}(\omega) = E_{z \sim \mathcal{D}} l(\omega, z)$

- initialize: $\omega^{(1)} = 0$
- for $i = 1, \dots, T$
 - ◆ choose $z_t \sim \mathbb{D}$
 - ◆ let $\mathbf{v}_t \in \partial l(\omega^{(t)}, z)$
 - ◆ update $\omega^{(t+1)} = \omega^{(t)} - \eta \mathbf{v}_t$
- output $\bar{\omega} = \frac{1}{T} \sum_{t=1}^T \omega^{(t)}$



Intuition.

- So far the learning is based on the empirical risk $L_S(\omega)$, how about directly minimizing $L_{\mathcal{D}}(\omega)$.
- Recall the update rule: $\omega^{(t+1)} = \omega^{(t)} - \eta \nabla L_{\mathcal{D}}(\omega^{(t)})$ where $\nabla L_{\mathcal{D}}(\omega^{(t)}) = E_{z \sim \mathcal{D}} \nabla l(\omega, z)$
- We can not calculate $\nabla L_{\mathcal{D}}(\omega^{(t)})$ because we don't know \mathcal{D} , but we can estimate it by $\nabla l(\omega, z)$ for $z \sim \mathcal{D}$.
- If we take a step in the direction $\mathbf{v} = \nabla l(\omega, z)$, then in expectation we are moving in the right direction. Namely \mathbf{v} is an unbiased estimate of the gradient.

□

(None)-4f977b3 (2021-10-09) – 16 / 30

Analysing Stochastic Gradient Descent for Lipschitz functions

Consider a convex-Lipschitz-bounded learning problem with parameters ρ , and B be an upper bound on $\|\omega^*\|$. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_{\mathcal{D}}(\omega)$ with a number of iterations (i.e., number of examples)

$T \geq \frac{B^2 \rho^2}{\epsilon^2}$ and with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output of SGD satisfies:

$$E[L_{\mathcal{D}}(\bar{\omega})] \leq \min_{\omega \in \mathcal{H}} L_{\mathcal{D}}(\omega) + \epsilon$$

Intuition.

1. Assuming the Lipschitz $\|\cdot\| \leq \rho$ for all t , from previous GD analysis we have

$$\sum_{t=1}^T \langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle \leq \frac{B^2}{2\eta} + \frac{\eta \rho^2 T}{2}$$

2. In particular, if $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, we have $\sum_{t=1}^T \langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle \leq B\rho\sqrt{T}$
3. Take expectation of both sides w.r.t. the randomness of choosing z_1, \dots, z_T we obtain: $E_{z_1, \dots, z_T} [\sum_{t=1}^T \langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle] \leq B\rho\sqrt{T}$

□

Intuition.


4. From the law of *total expectation* $E_{\alpha}[g(\alpha)] = E_{\beta} E_{\alpha}[g(\alpha)|\beta]$, we have:

$$E_{z_1, \dots, z_T} [\langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle] = E_{z_1, \dots, z_{t-1}} E_{z_t} [\langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle | z_1, \dots, z_{t-1}] \leq B\rho\sqrt{T}$$

5. Once we know z_1, \dots, z_{t-1} and the value of $\omega^{(t)}$ is not random, therefore $E_{z_1, \dots, z_T} [\langle \mathbf{v}_t, \omega^{(t)} - \omega^* \rangle | z_1, \dots, z_{t-1}] = \langle E_{z_t} \nabla l(\omega^{(t)}, z_t), \omega^{(t)} - \omega^* \rangle = \langle \nabla L_{\mathcal{D}}(\omega^{(t)}), \omega^{(t)} - \omega^* \rangle$
6. We then have $E_{z_1, \dots, z_T} [\sum_{t=1}^T \langle \nabla L_{\mathcal{D}}(\omega^{(t)}), \omega^{(t)} - \omega^* \rangle] \leq B\rho\sqrt{T}$
7. By convexity, $E_{z_1, \dots, z_T} [\sum_{t=1}^T L_{\mathcal{D}}(\omega^{(t)}) - L_{\mathcal{D}}(\omega^*)] \leq B\rho\sqrt{T}$
8. Divided by T and using convexity again, we have $E_{z_1, \dots, z_T} [L_{\mathcal{D}}(\frac{\sum_{t=1}^T \omega^{(t)}}{T})] \leq L_{\mathcal{D}}(\omega^*) + \frac{B\rho}{\sqrt{T}}$

□

Newton's method: second-order method



To minimize a twice differentiable convex function $f(\omega)$, assume the gradient of $f : \mathcal{R}^d \mapsto \mathcal{R}$ at ω , denoted as $g = \nabla f(\omega) = (\frac{\partial f(\omega)}{\partial \omega_1}, \dots, \frac{\partial f(\omega)}{\partial \omega_d})$ is the **gradient** of f at ω , and the Hessian $H = \nabla^2 f(\omega)$

- Start with initial $\omega^{(1)}$ (usually, the zero vector)
- At iteration t , update $\omega^{(t+1)} = \omega^{(t)} - H^{-1}g$

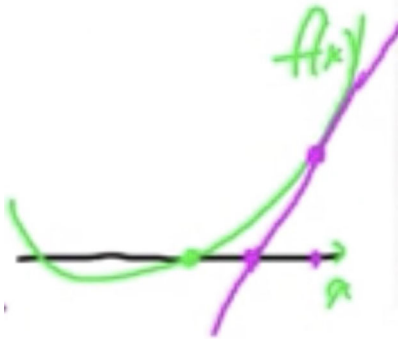
Analogy (1D).

- For zero finding in $f(\omega)$, Newton's method iterates as:

$$\omega^{(t+1)} = \omega^{(t)} - \frac{f(\omega^{(t)})}{f'(\omega^{(t)})}$$

- Finding the minimum or maximum of f is equal to zero finding in function f' , hence Newton's method of optimisation iterates as

$$\omega^{(t+1)} = \omega^{(t)} - \frac{f'(\omega^{(t)})}{f''(\omega^{(t)})}$$



□

Intuition.

- Let f to be sufficiently smooth. By Taylor's approximation, if close to $\omega^{(t)}$, we have

$$\begin{aligned} f(x) &\cong f(\omega^{(t)}) + g^T(x - \omega^{(t)}) + \frac{1}{2}(x - \omega^{(t)})^T H(x - \omega^{(t)}) = f(\omega^{(t)}) + g^T(x - \omega^{(t)}) + \frac{1}{2}(x^T Hx - 2\omega^{(t)T} Hx + \omega^{(t)T} H\omega^{(t)}) \\ &= \frac{1}{2}x^T Hx + (g - H\omega^{(t)})^T x + C \end{aligned}$$

- Take the first derivative, we have $\nabla f = Hx + (g - H\omega^{(t)}) = 0$. So we have $x = -H^{-1}(g - H\omega^{(t)}) = \omega^{(t)} - H^{-1}g$

□

Properties.

- Because $\nabla^2 f = H$, so f is minimized when H is positive self-definite.
- H may fail to be positive definite, rather than inverse H , we may solve $Hy = g$ for y , then use $\omega^{(t+1)} = \omega^{(t)} - y$
- We may also introduce the learning rate η , hence $\omega^{(t+1)} = \omega^{(t)} - \eta y$


□

(None)-4f977b3 (2021-10-09) – 18 / 30

Convex Learning Problems

19 / 30

Convex Learning Problems



A learning problem $(\mathcal{H}, \mathcal{Z}, l)$ is called **convex** if the hypothesis class \mathcal{H} is a convex set and for all $z \in \mathcal{Z}$, the loss function $l(\cdot, z)$ is a convex function, where for any z , $l(\cdot, z)$ denotes the function $f : \mathcal{H} \mapsto \mathcal{R}$ defined by $f(\omega) = l(\omega, z)$.

Claims.

- $ERM_{\mathcal{H}}$ w.r.t. a convex learning problem is a convex optimization

$$\min_{\omega \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m l(\omega, z_i)$$

Example Least squares: $\mathcal{H} = \mathcal{R}$, $\mathcal{Z} = \mathcal{R}^d \times \mathcal{R}$, $l(\omega, (x, y)) = (\langle \omega, x \rangle - y)^2$

- Implementing the ERM rule for convex learning problems can be done efficiently, but is convexity a sufficient condition for the learnability of a problem?
- Not all convex learning problems over \mathcal{R}^d are learnable.
 - ◆ The intuitive reason is numerical stability.
- With two additional mild conditions, we obtain learnability.

□

(None)-4f977b3 (2021-10-09) – 20 / 30

Convex-Lipschitz-bounded learning problem

A learning problem $(\mathcal{H}, \mathcal{Z}, l)$ is called **convex-Lipschitz-Bounded**, with parameters ρ and B if the following holds:

- The hypothesis class \mathcal{H} is a convex set, and $\forall \omega \in \mathcal{H}$ we have $\|\omega\| \leq B$.
- For all $z \in \mathcal{Z}$, the loss function, $l(\cdot, z)$, is a convex and ρ -Lipschitz function.

Examples.

- $\mathcal{H} = \{\omega \in \mathbb{R}^d : \|\omega\| \leq B\}$
- $X = \{x \in \mathbb{R}^d : \|x\| \leq \rho\}$ and $Y = \mathbb{R}$
- $l(\omega, (x, y)) = |\langle \omega, x \rangle - y|$

□

Convex-Smooth-bounded learning problem

A learning problem $(\mathcal{H}, \mathcal{Z}, l)$ is called **Convex-Smooth-Bounded**, with parameters β and B if the following holds:

- The hypothesis class \mathcal{H} is a convex set, and $\forall \omega \in \mathcal{H}$ we have $\|\omega\| \leq B$.
- For all $z \in \mathcal{Z}$, the loss function, $l(\cdot, z)$, is convex, non-negative and β -smooth.

Statements.

- A function f is β -smooth if it is differentiable and its gradient is β -Lipschitz.
- We require that the loss function to be non-negative, this is to ensure that the loss function is self-bounded.

□

Examples.

- $\mathcal{H} = \{\omega \in \mathbb{R}^d : \|\omega\| \leq B\}$
- $X = \{x \in \mathbb{R}^d : \|x\| \leq \beta/2\}$ and $Y = \mathbb{R}$
- $l(\omega, (x, y)) = (\langle \omega, x \rangle - y)^2$

□

Surrogate Loss Functions

For problems with non-convex loss functions, one popular approach is to upper bound the non-convex loss function using a convex **surrogate loss function**:

- It should be convex
- It should upper bound the original loss

Notes.

- When trying to minimize the empirical risk with respect to a non convex loss function, we might encounter local minima.
- Also solving the ERM problem with respect to the 0 – 1 loss in the unrealisable case is known to be NP-hard.

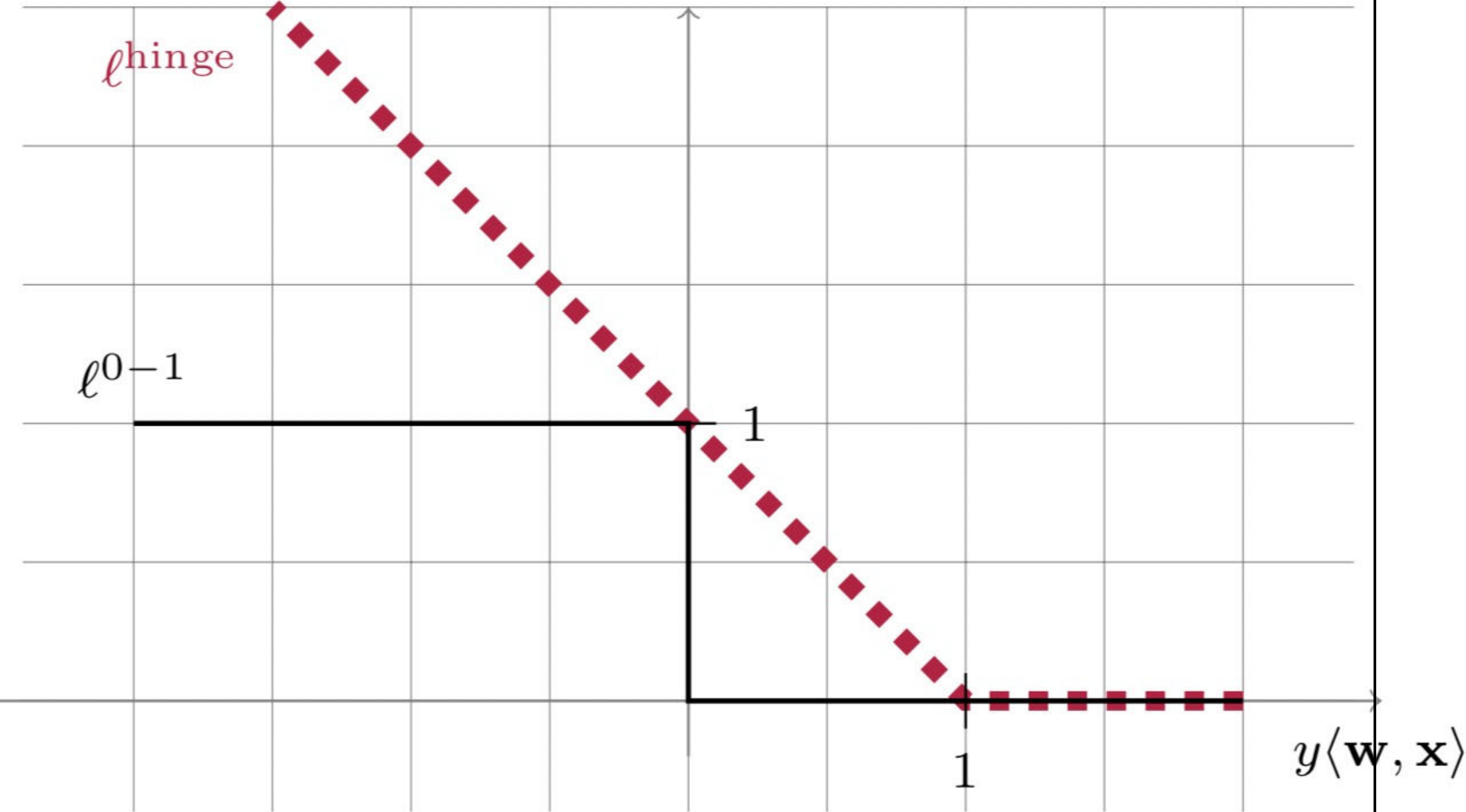
The 0 – 1 Loss Function and Hinge Loss

In the context of learning halfspaces, the 0 – 1 loss is not convex:

$$l^{0-1}(\omega, (x, y)) = 1_{y \neq \text{sign}(\langle \omega, x \rangle)} = 1_{y \langle \omega, x \rangle \leq 0}$$

👍 We can define a convex surrogate for it, the hinge loss:

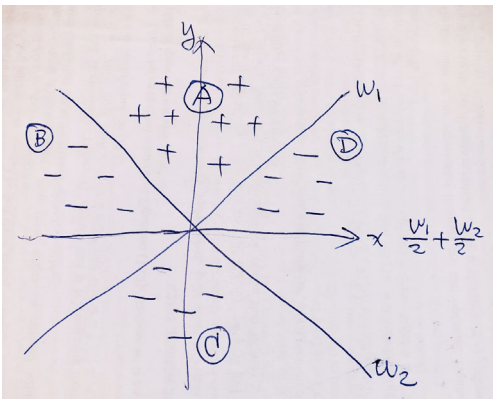
$$l^{\text{hinge}}(\omega, (x, y)) = \max\{0, 1 - y \langle \omega, x \rangle\}$$



Why $l^{0-1}(\omega, (x, y))$ non-convex?

- Imagine a special case as in the right figure. We have $l^{0-1}(\omega_1, (x, y)) = \frac{B}{m}$ and $l^{0-1}(\omega_2, (x, y)) = \frac{D}{m}$.
- If taking the average of ω_1 and ω_2 , we have $\omega_a = \frac{1}{2}\omega_1 + \frac{1}{2}\omega_2$
- $l^{0-1}(\omega_a, (x, y)) = \frac{B+D}{m} > \frac{1}{2}l^{0-1}(\omega_1, (x, y)) + \frac{1}{2}l^{0-1}(\omega_2, (x, y))$.
- If l^{0-1} is convex, then we should have $l^{0-1}(\frac{1}{2}\omega_1 + \frac{1}{2}\omega_2, (x, y)) \leq \frac{1}{2}l^{0-1}(\omega_1, (x, y)) + \frac{1}{2}l^{0-1}(\omega_2, (x, y))$. However, the above shows that this is not true.
- Hence it is non-convex.

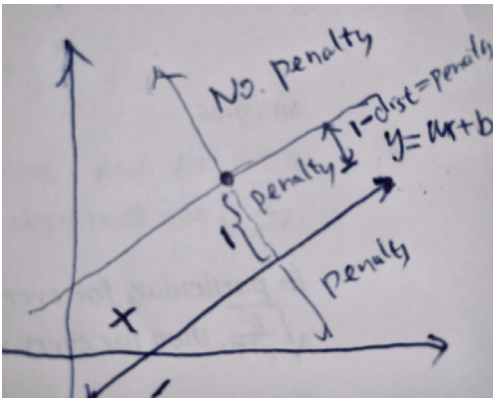
□



Why $l^{\text{hinge}}(\omega, (x, y))$ works?

- Both \max and $1 - y \langle \omega, x \rangle$ are convex, so l^{hinge} is convex.
- The distance from point x to hyperplane defined by $\omega = (\omega_x, \omega_0)$ where $\|\omega\| = 1$ is $|\langle \omega, x \rangle| = |\langle \omega_x, x \rangle + \omega_0|$.
 - ◆ Projecting x to the hyperplane and assume the point to be: $\vec{v} = \vec{x} - (\langle \omega_x, \vec{x} \rangle + \omega_0)\omega_x$
 - ◆ We can see $\langle \omega, x \rangle = \langle \omega_x, \vec{v} \rangle + \omega_0 = \langle \omega_x, \vec{x} \rangle - (\langle \omega_x, \vec{x} \rangle + \omega_0)\|\omega_x\|^2 + \omega_0 = 0$
- Geometrically, l^{hinge} represents the penalty or loss which considers the margin of the classifier:
 - ◆ No penalty for corrected examples larger than the margin
 - ◆ $1 - y \langle \omega, x \rangle$ for corrected examples within the margin, or wrong examples

□

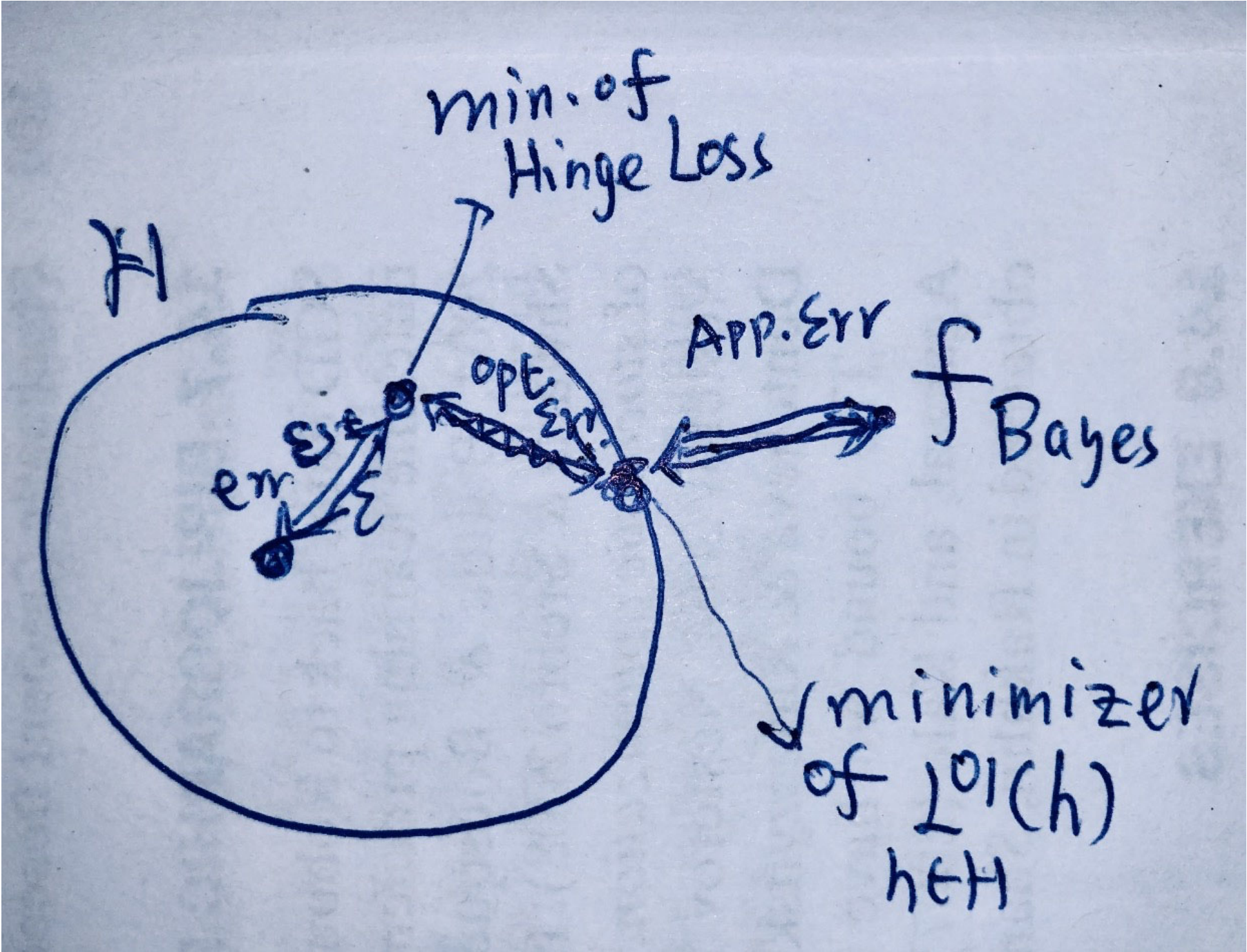


Error Decomposition Revisited

Suppose we have a learner for the hinge-loss that guarantees: $L_{\mathcal{D}}^{hinge}(A(S)) \leq \min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{hinge}(\omega) + \epsilon$. Using the surrogate, we have $L_{\mathcal{D}}^{0-1}(A(S)) \leq \min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{hinge}(\omega) + \epsilon$. We can further rewrite the upper bound as:

👍

$$L_{\mathcal{D}}^{0-1}(A(S)) \leq \min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{0-1}(\omega) + (\min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{hinge}(\omega) - \min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{0-1}(\omega)) + \epsilon$$



Notes.

Approximation Error $L_{\mathcal{D}}^{0-1}(\omega)$ which measures how well the hypothesis class performs on the distribution.

Estimation Error This is the error that results from the fact that we only receive a training set and do not observe the distribution \mathcal{D}

Optimisation Error $(\min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{hinge}(\omega) - \min_{\omega \in \mathbb{H}} L_{\mathcal{D}}^{0-1}(\omega))$ which measures the difference between the approximation error with respect to the surrogate loss and the approximation error with respect to the original loss.

- It is the result of our inability to minimize the training loss with respect to the original loss.
- Its size depends on the specific data distribution and on the specific surrogate loss.

□

(None)-4f977b3 (2021-10-09) – 26 / 30

Quiz

27 / 30

SGD with Projection Step

?

In the SGD or GD, the norm of ω is required to be at most B . But there is no guarantee that ω in SGD/GD will satisfy it.

- We can resolve this issue by a projection step: we first subtract a subgradient from the current value of ω , and then project the resulting vector onto the hypothesis class with norm at most B .
 1. $\omega^{t+\frac{1}{2}} = \omega^t - \eta v_t$
 2. $\omega^{t+1} = \arg\max_{\omega \in \mathcal{H}} |\omega - \omega^{t+\frac{1}{2}}|$

The project step replaces the current value of ω by the vector in \mathcal{H} closest to it.

- Projection Lemma: Let \mathcal{H} be a closed convex set, and let v be the projection of ω onto \mathcal{H} , namely, $v = \arg\min_{x \in \mathcal{H}} |x - \omega|^2$. Then prove that: for every $u \in \mathcal{H}$, $|\omega - u|^2 - |v - u|^2 \geq 0$.




(None)-4f977b3 (2021-10-09) – 28 / 30

Questions?

Contact Information

Associate Professor **GANG LI**
School of Information Technology
Deakin University
Geelong, Victoria 3216, Australia



-  GANGLI@TULIP.ORG.AU
-  OPEN RESOURCES OF TULIP-LAB
-  TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING