COS234:

Introduction To Machine Learning

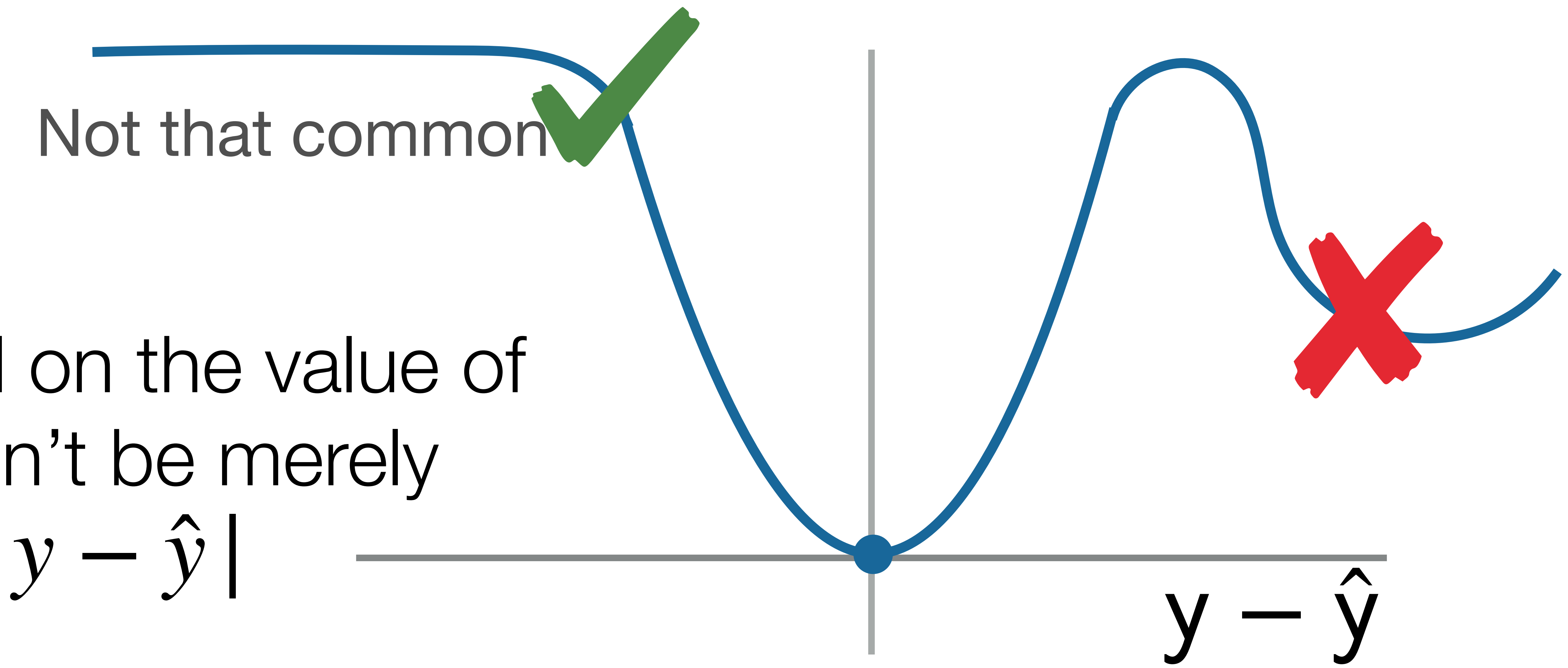
Prof. Yoram Singer

PRINCETON UNIVERSITY

Topic: Linear Regression

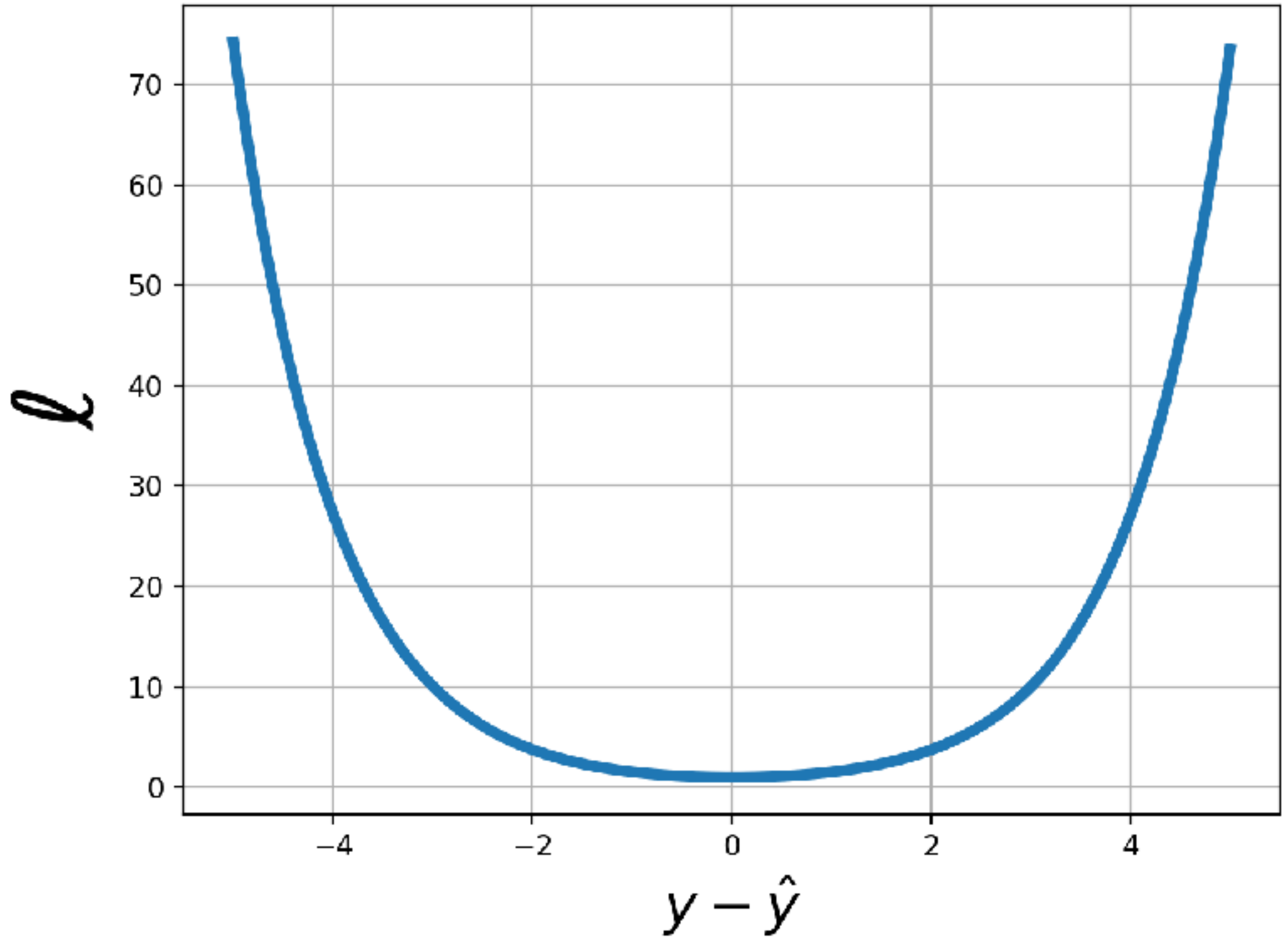
Properties of ℓ

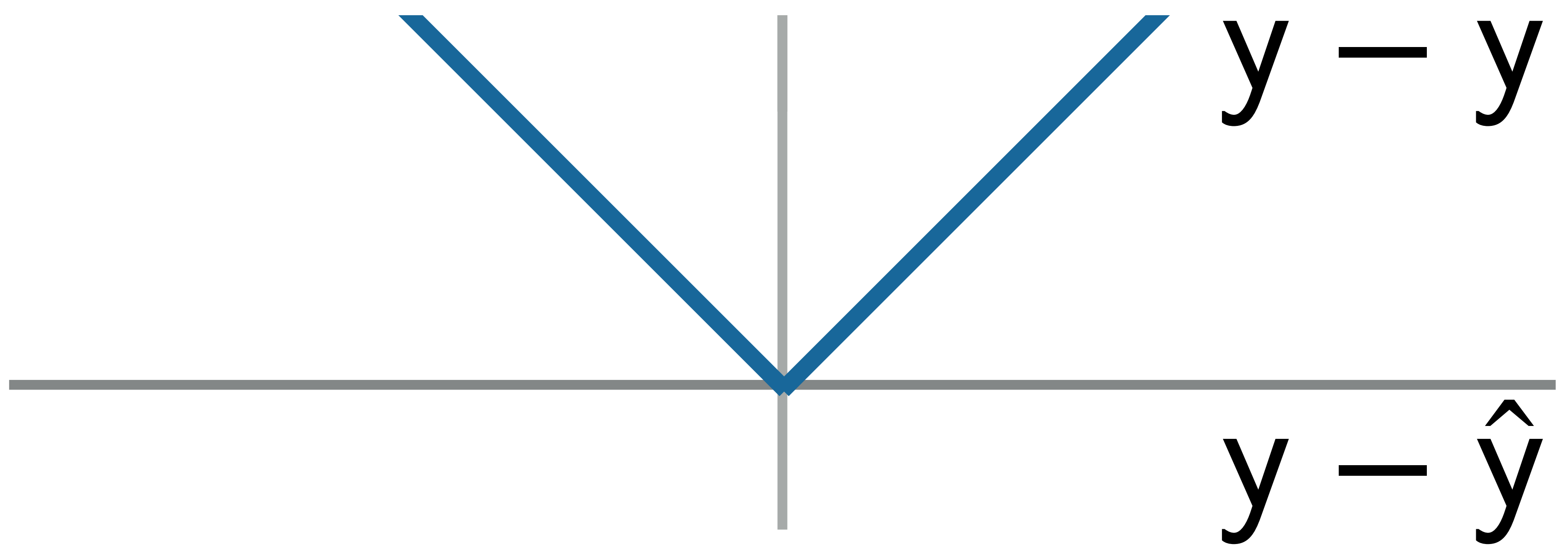
- When $y = \hat{y}$ loss should be 0
- When $y \neq \hat{y}$ loss should be ≥ 0
- If $|y_2 - \hat{y}_2| > |y_1 - \hat{y}_1|$ we typically want $\ell(y_2, \hat{y}_2) > \ell(y_1, \hat{y}_1)$



- May depend on the value of y and wouldn't be merely function of $|y - \hat{y}|$

Regression Losses





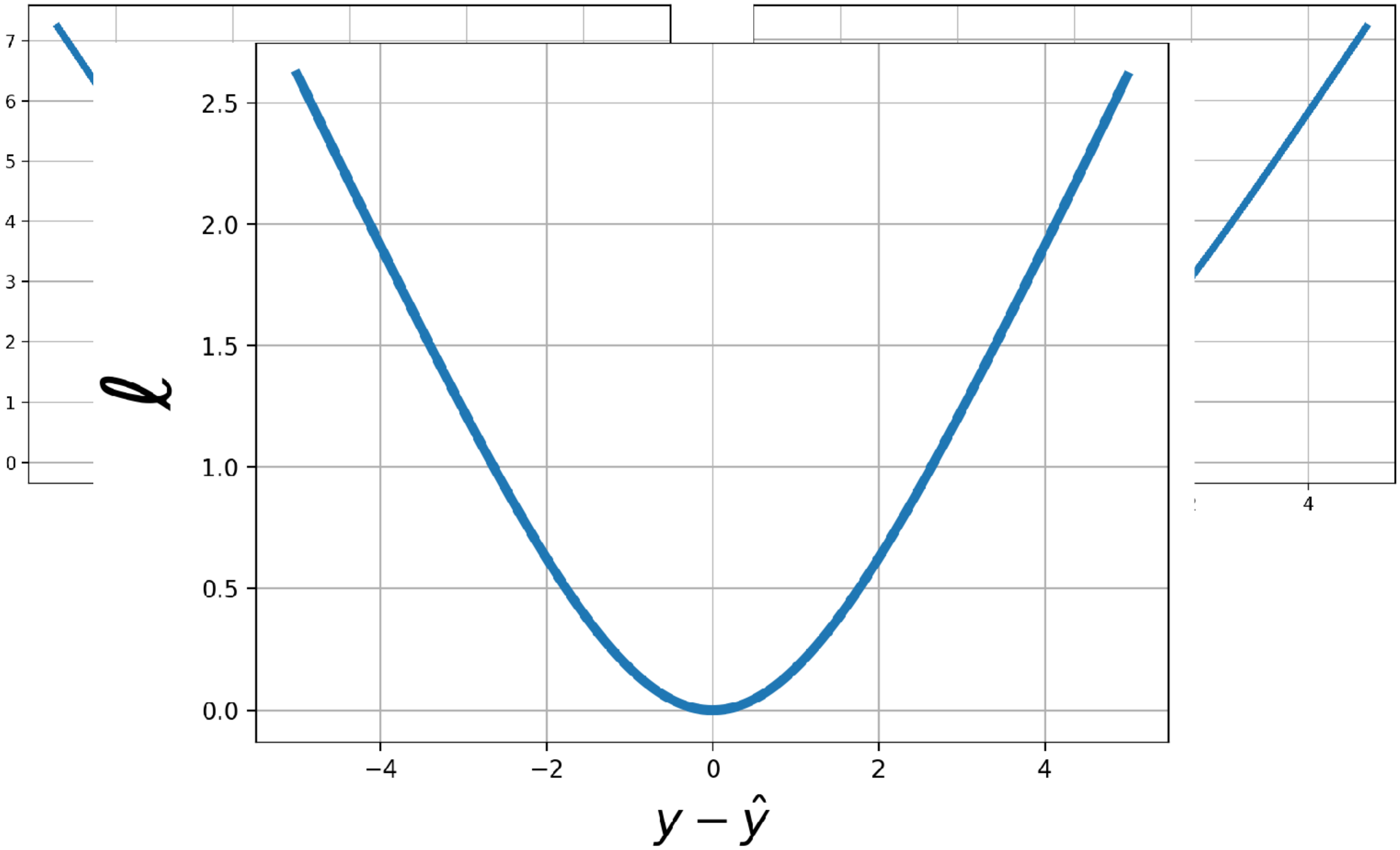
- Squared loss $\ell(y, \hat{y}) = (y - \hat{y})^2$
- Absolute loss $\ell(y, \hat{y}) = |y - \hat{y}|$
- Exponential loss $\ell(y, \hat{y}) = e^{y-\hat{y}} + e^{\hat{y}-y}$

Symmetrization of Losses

- Given $f : \Re \to \Re$ bounded from below: $\exists c : \; f(z) > c$
- Symmetrization at 0:

$$\ell(z) = \frac{1}{2}\big(f(z) + f(-z)\big) - f(0) \;\;\Rightarrow\;\; \ell(0) = 0$$

- Use $z = y - \hat{y}$
- Exp-Loss: $f(z) = e^{z}$
- Log-Loss: $f(z) = \log\left(1 + e^{z}\right)$



ℓ
2.5
2.0
1.5
1.0
0.5
0.0
−4
−2
0
2
4
$y - \hat{y}$

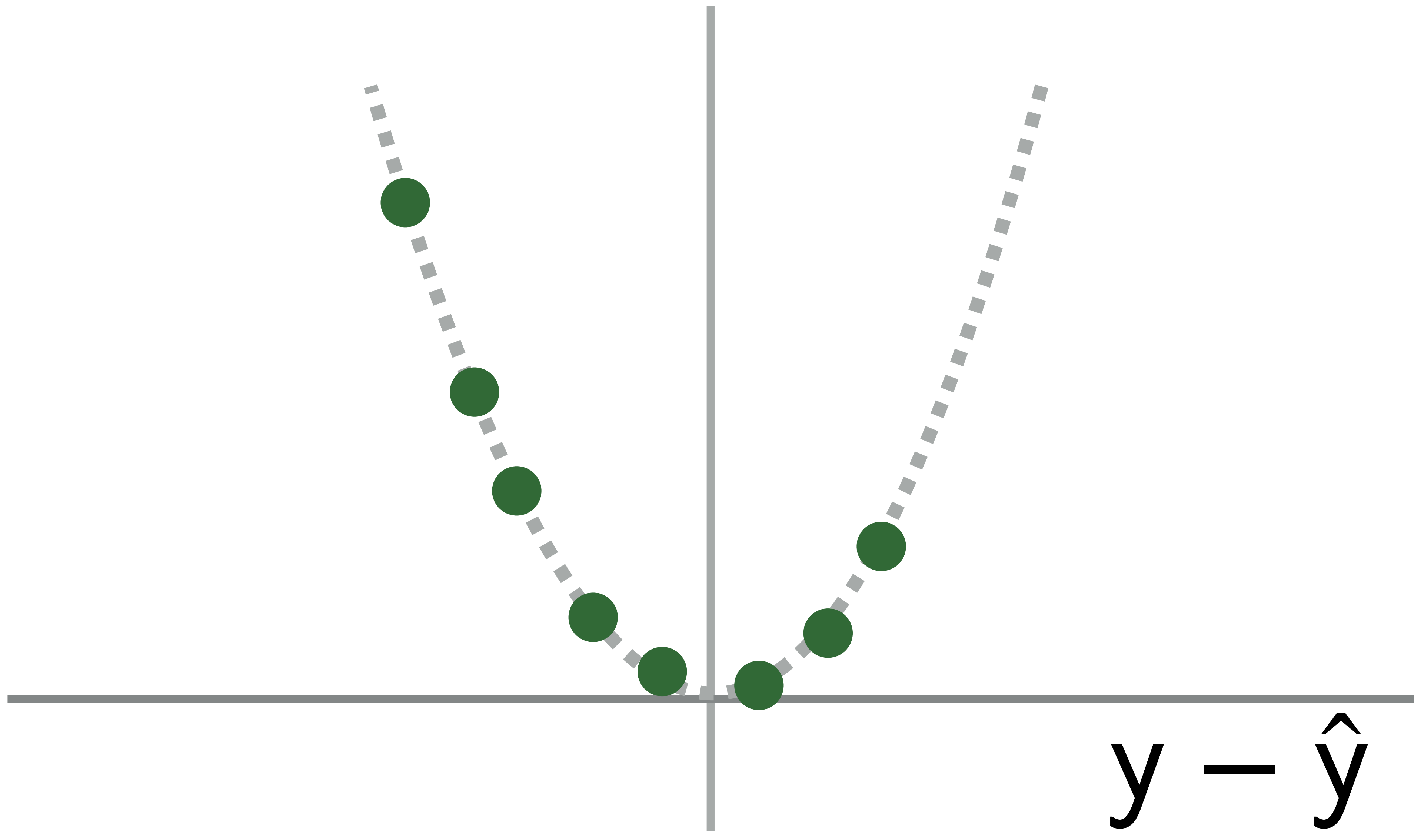
Training Loss

Average (why average?) loss over all examples

$$\frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i, \hat{y}_i(\mathbf{w})\right)$$

For squared loss we can write

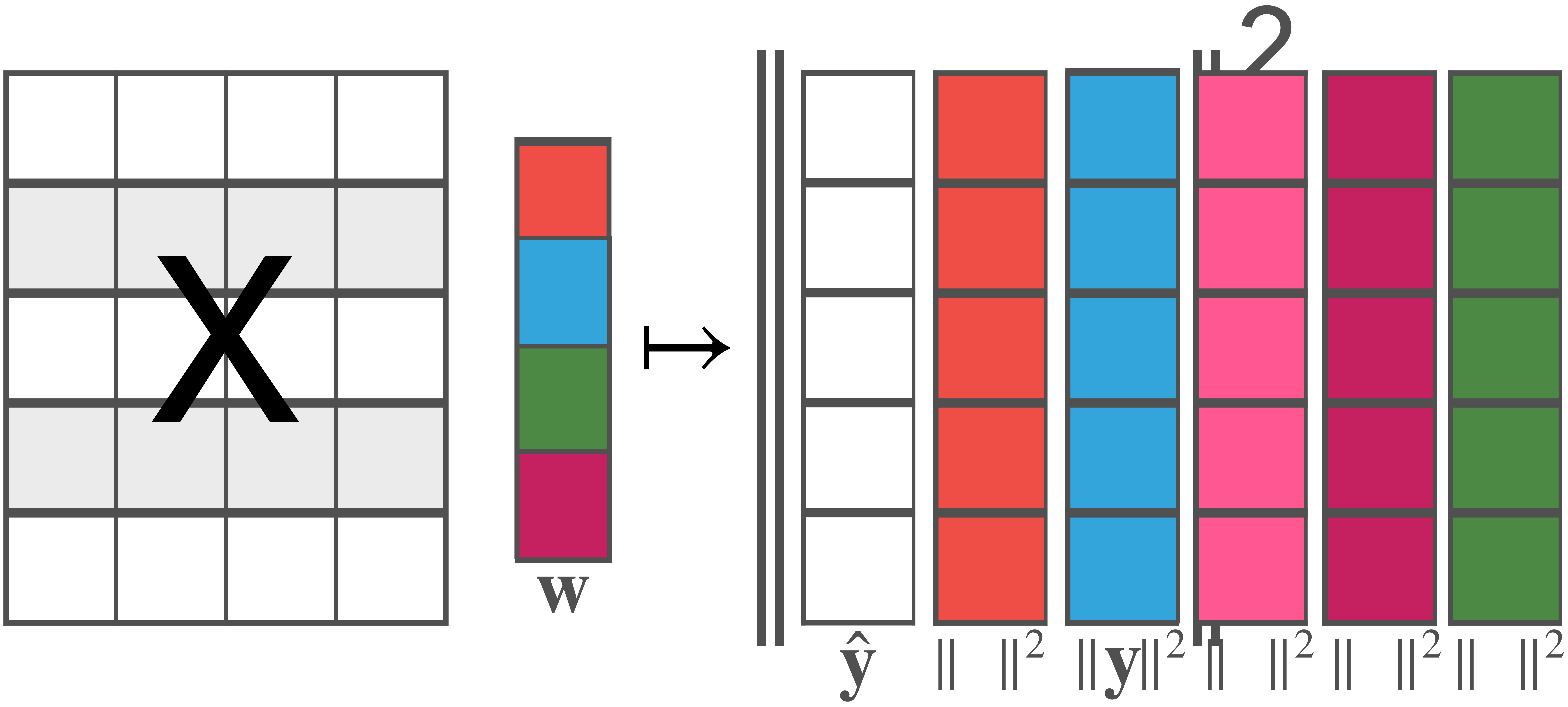
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\|\mathbf{y} - \hat{\mathbf{y}}\|^2$$



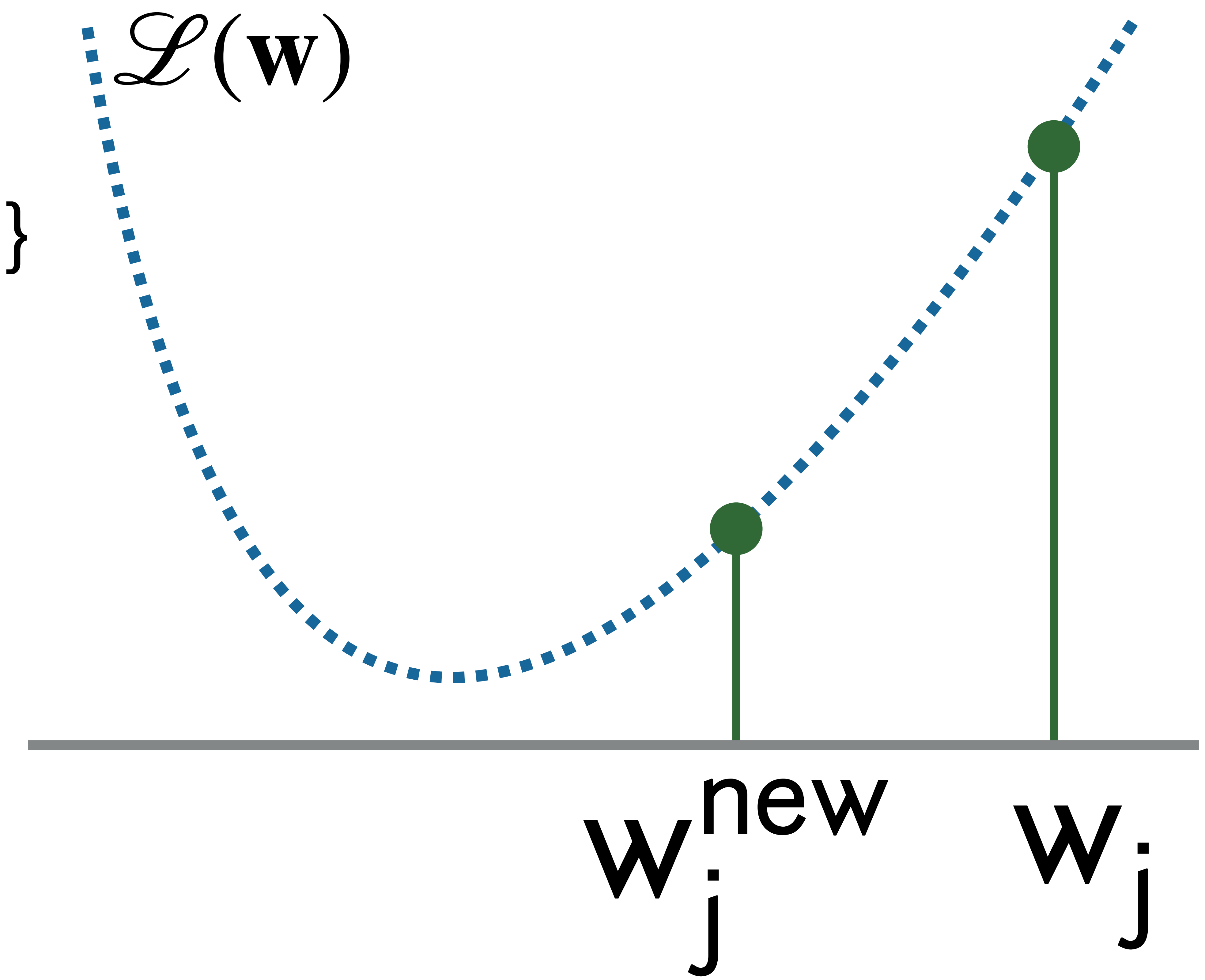
Least Squares Regression

- Matrix $\mathbf{X}$ of $\mathbf{n}$ examples in $\mathbf{d}$ dimensions
- Targets are real-valued: $\mathbf{y}$ is an $\mathbf{n}$ dimensional vector
- Examples for this setting?
- Squared loss: $\ell(\mathrm{y}, \hat{\mathrm{y}}) = (\mathrm{y} - \hat{\mathrm{y}})^2$
- Find w such that training loss is as small as possible:

$$\boxed{\mathscr{L}(\mathbf{w})} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i(\mathbf{w})\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}\cdot\mathbf{x_i} - y_i\right)^2$$


X
w
ŷ
‖y‖²
2

- We want to find $\mathbf{w}$ so as to (approximately) minimize $\mathscr{L}(\mathbf{w})$
- Assume that we are given an initial guess $\mathbf{w}$
  [in practice initialize $\mathbf{w} = \mathbf{0} = (0,0, \ldots ,0)$]
- Loop:
  - Pick an index j at random from $\{1,2, \ldots ,d\}$
  - Replace w_j with a <u>better</u> estimate w_j^{new}
- Until when ?



$\mathrm{w}_j \mapsto \mathrm{w}_j^{\mathrm{new}}$

1. Define error vector $\mathbf{e} = \hat{\mathbf{y}} - \mathbf{y}$ and write loss as $\dfrac{1}{n}||\mathbf{e}||^2$
2. Define $a = w_j^{\text{new}} - w_j$
3. Define $\mathscr{L}(a)$ $\left[\text{overloading } \mathscr{L}(a), \mathscr{L}(\mathbf{w})\right]$ then
$$\mathscr{L}(a) = \frac{1}{n}\sum_{i=1}^{n}\left(e_i + aX_{ij}\right)^2$$
4. Because
$$e_i^{\text{new}} = \hat{y}_i^{\text{new}} - y_i = \mathbf{w}^{\text{new}} \cdot \mathbf{x_i} - y_i = \mathbf{w} \cdot \mathbf{x_i} - y_i + aX_{ij} = e_i + aX_{ij}$$
5. What value should we choose for a ?

Chain Rule

If $f(a) = h\big(\overbrace{r(a)}^{\equiv z}\big)$ then $\dfrac{df(a)}{da} = \dfrac{dh(z)}{dz}\Big|_{z=r(a)} \dfrac{dr(a)}{da}$

Example: assume $f(a) = \big(\log(a)\big)^2$

Define $h(z)=z^2$ and $r(a)=\log(a)$ then $\dfrac{dh}{dz} = 2z$ and $\dfrac{dr}{da} = \dfrac{1}{a}$

Then $\dfrac{df(a)}{da} = 2z\dfrac{1}{a}$ where $z = \log(a)$

We get $\dfrac{df(a)}{da} = \dfrac{2\log(a)}{a}$

Function Minimization

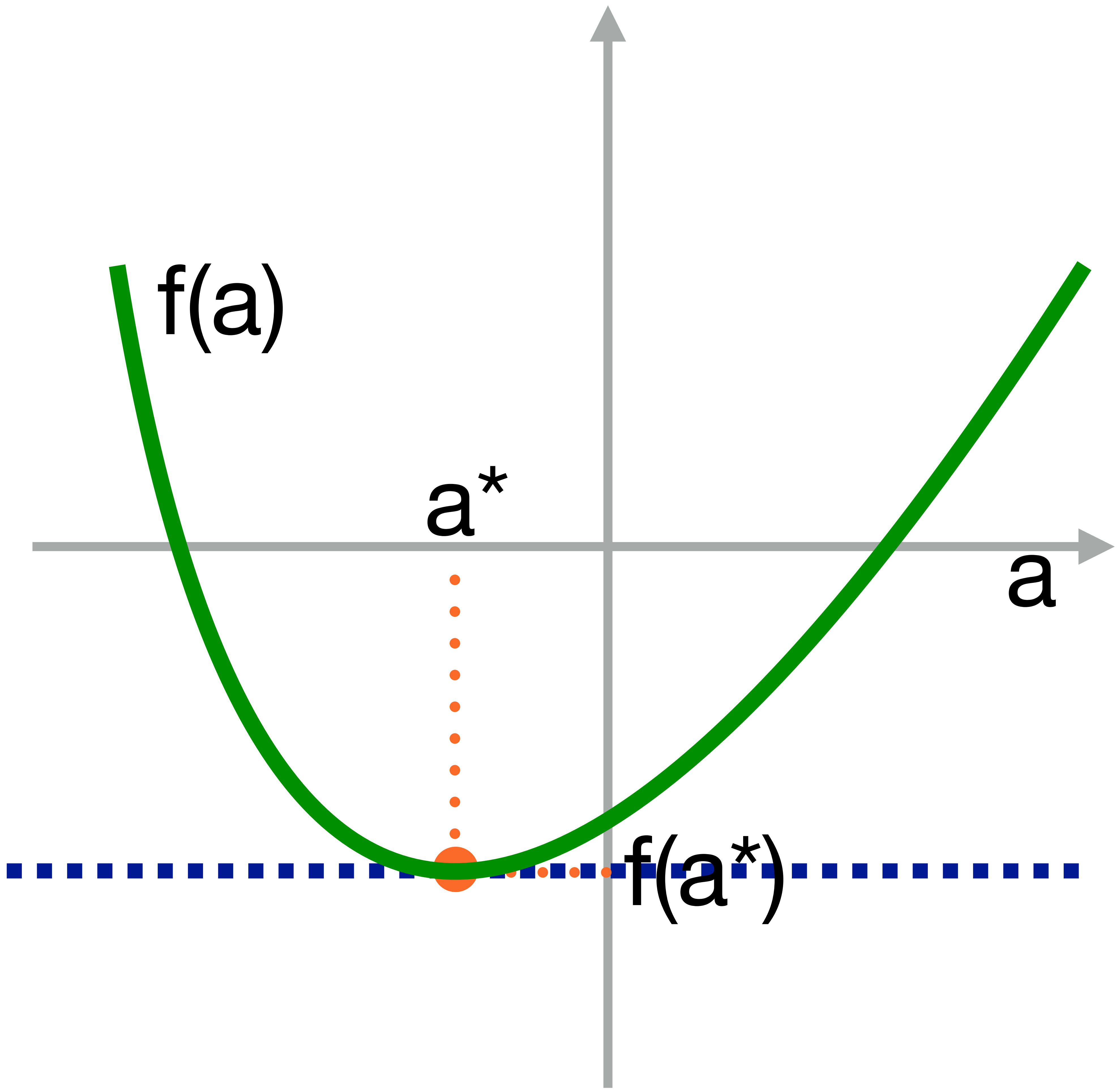
- If a^* is the minimum point of f then

$$f(a^*) < f(b) \text{ for all } b \neq a^*$$

- Derivate of $f(a)$ at a^* is zero

$$\left.\frac{df}{da}\right|_{a=a^*} \equiv f'(a^*) = 0$$

- To find the minimum of f solve $\dfrac{df}{da} = 0$



Back to $w_j \mapsto w_j^{new}$

1. We need to take the derivate of $\mathcal{L}(a) = \frac{1}{n}\sum_{i=1}^{n}\left(e_i + aX_{ij}\right)^2$

2. Let $z_i = e_i + aX_{ij}$ for which $\frac{dz_i}{da} = X_{ij}$ and write $\mathcal{L}(a) = \frac{1}{n}\sum_{i=1}^{n} z_i^2$

$$\frac{d\mathcal{L}}{da} = \frac{1}{n}\sum_{i=1}^{n}\frac{dz_i^2}{dz_i}\frac{dz_i}{da} = \frac{1}{n}\sum_{i=1}^{n} 2z_iX_{ij} = \frac{2}{n}\sum_{i=1}^{n}\left(e_i + aX_{ij}\right)X_{ij} \text{ should be 0}$$

3. Therefore

$$-\sum_{i=1}^{n} e_i X_{ij} = \sum_{i=1}^{n} a X_{ij} X_{ij} \quad \Rightarrow \quad -\mathbf{e}\cdot X_{*j} = a\|X_{*j}\|^2 \quad \Rightarrow \quad a = -\frac{\mathbf{e}\cdot X_{*j}}{\|X_{*j}\|^2}$$

- Initialize: $\mathbf{w} = \mathbf{0} = (0,0, \ldots ,0)$

  Initialize: $\mathbf{e} = \hat{\mathbf{y}} - \mathbf{y} = -\mathbf{y}$
- Loop:
  - Pick j at random from $\{1,2, \ldots ,d\}$
  - Calculate

$$a = -\frac{\mathbf{e} \cdot \mathbf{X}_{*\mathbf{j}}}{\|\mathbf{X}_{*\mathbf{j}}\|^2}$$

  - Update $\mathbf{e} \leftarrow \mathbf{e} + a\,\mathbf{X}_{*\mathbf{j}}$
  - Update $w_j \leftarrow w_j + a$

Convergence

- Optimum (although we do not know it)

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{w} \cdot \mathbf{x_i} - y_i\right)^2$$

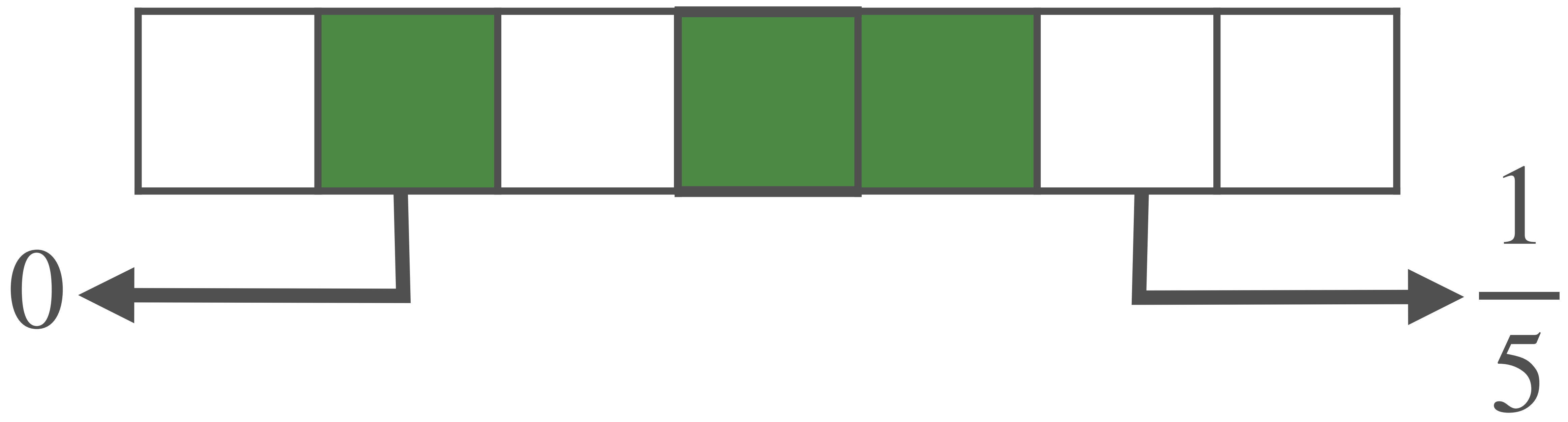
  which btw may not be unique (why?)

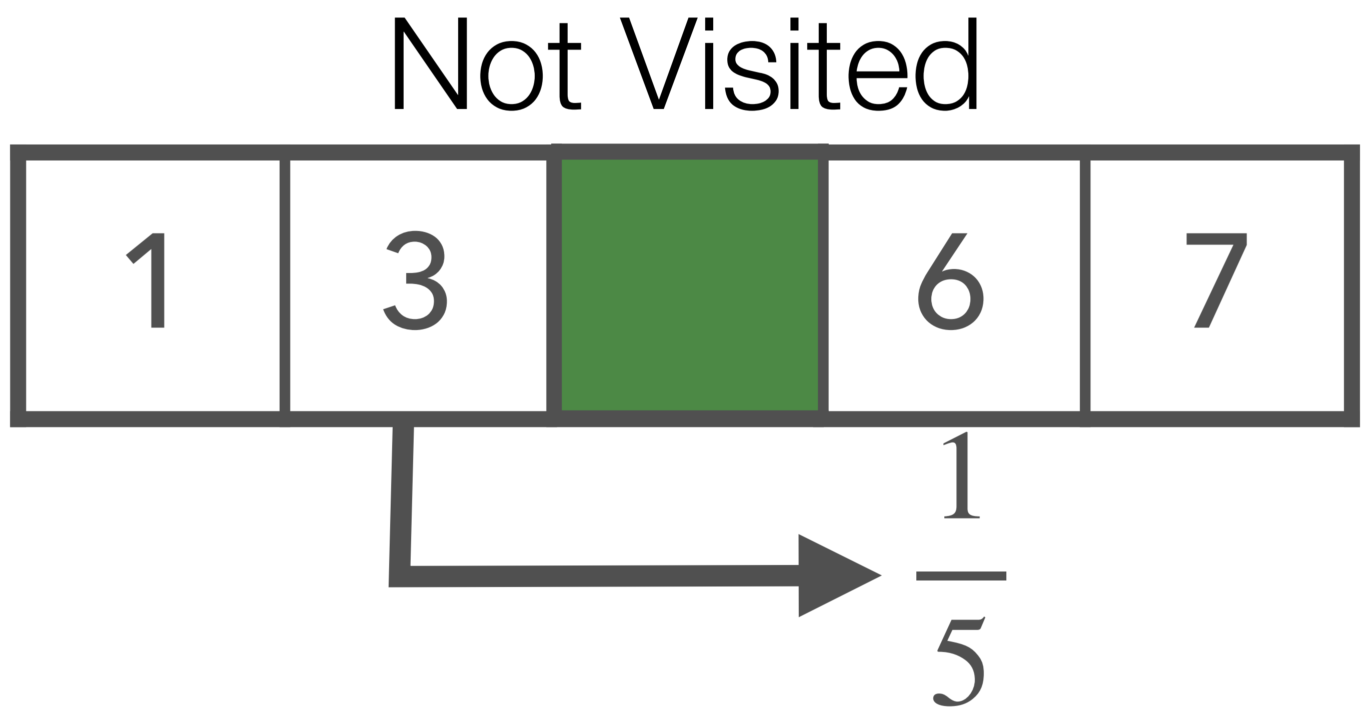
- Sequence algorithms generates: $\mathbf{w^1}, \mathbf{w^2}, \ldots, \mathbf{w^t}, \mathbf{w^{t+1}}, \ldots$
- Unless at optimum $\forall t:\ \mathscr{L}\left(\mathbf{w}^{\mathrm{t}}\right) > \mathscr{L}\left(\mathbf{w}^{\mathrm{t+1}}\right)$
- Under (mild) conditions $\mathscr{L}\left(\mathbf{w}^{\mathrm{t}}\right) \to \mathscr{L}\left(\mathbf{w}^\star\right)$ as $t \to \infty$

Cycling Through Variables

I. Random selection with replacement:

II. Random selection without replacement:





Termination Condition

I. Until you run out of time:

For t = 1, 2, … , T

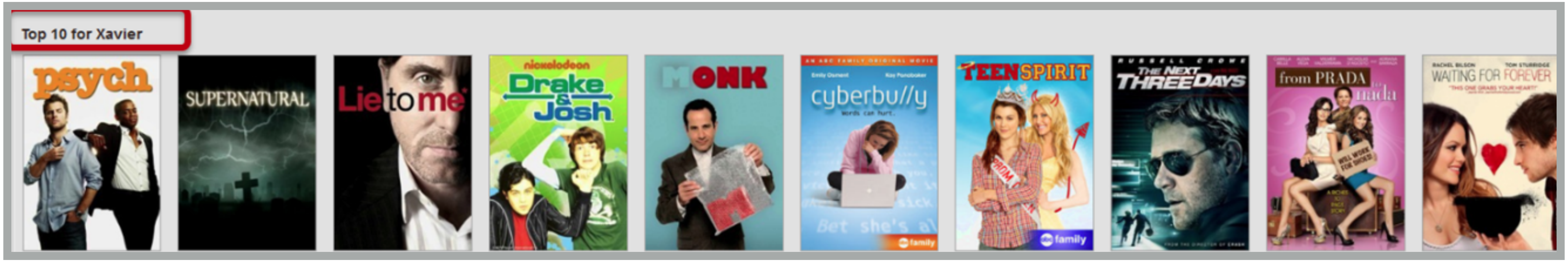
Update …

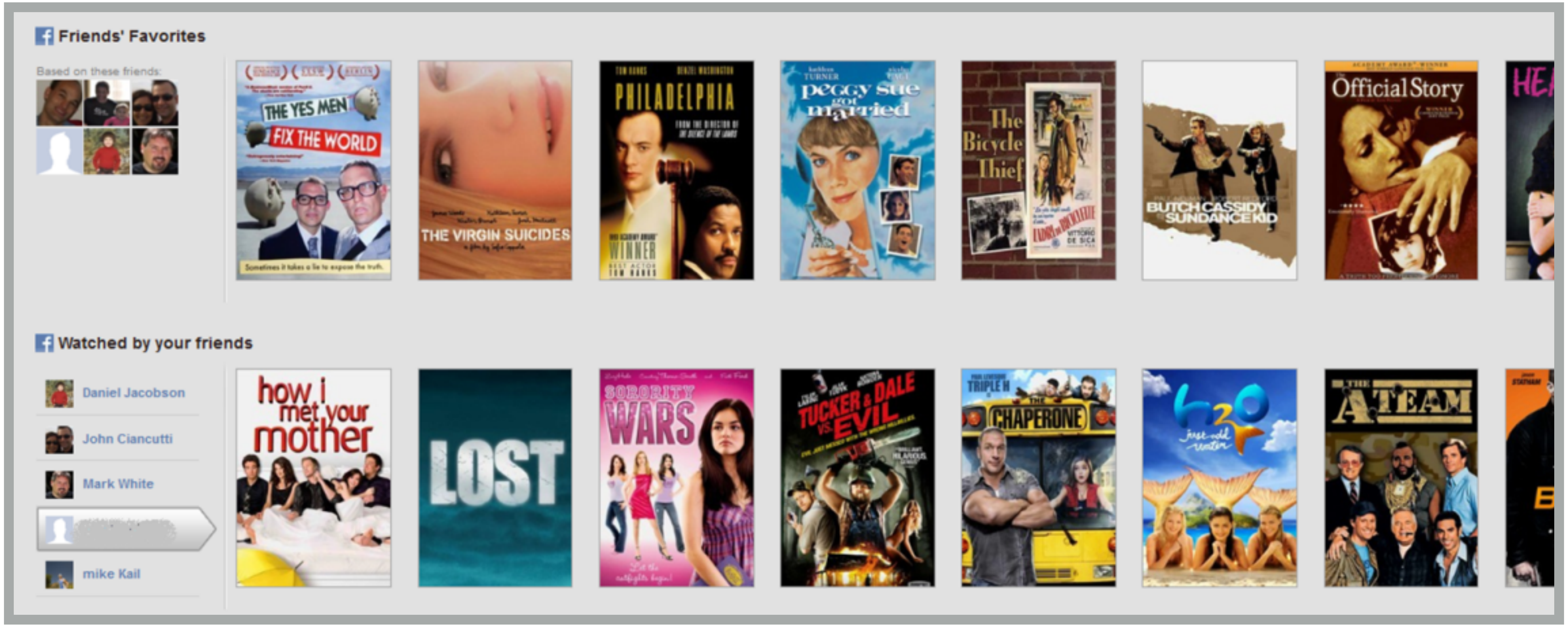
II. Until sufficient accuracy ϵ (for regression ~ in [0.001,0.1]) is established:

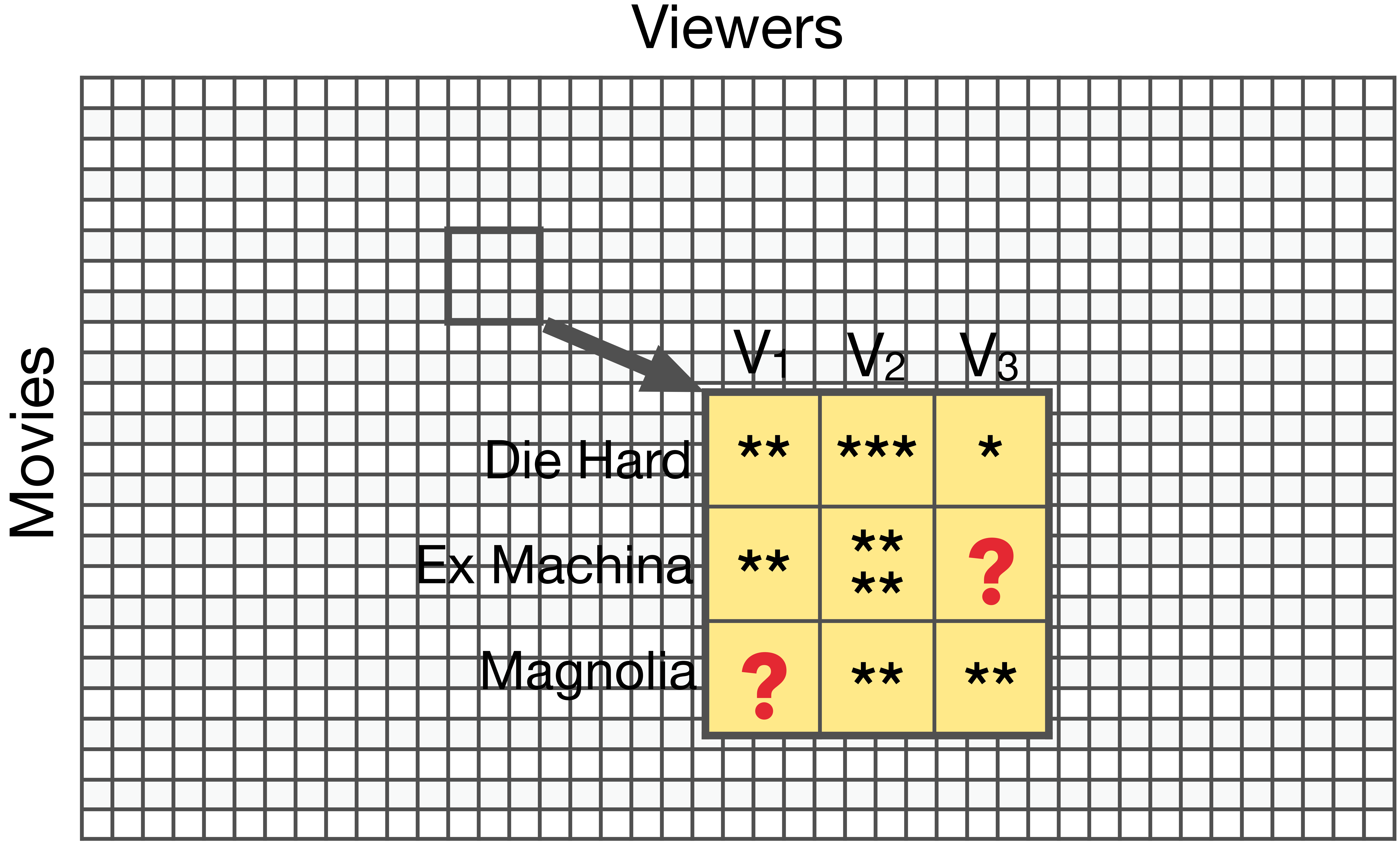
I. In objective: $$\frac{\mathscr{L}(\mathbf{w}^t) - \mathscr{L}(\mathbf{w}^{t+1})}{\mathscr{L}(\mathbf{w}^{t+1})} \leq \epsilon$$

II. In parameters: $$\frac{\|\mathbf{w}^t - \mathbf{w}^{t+1}\|}{\|\mathbf{w}^{t+1}\|} \leq \epsilon$$

The Netflix Prize Story


Top 10 for Xavier

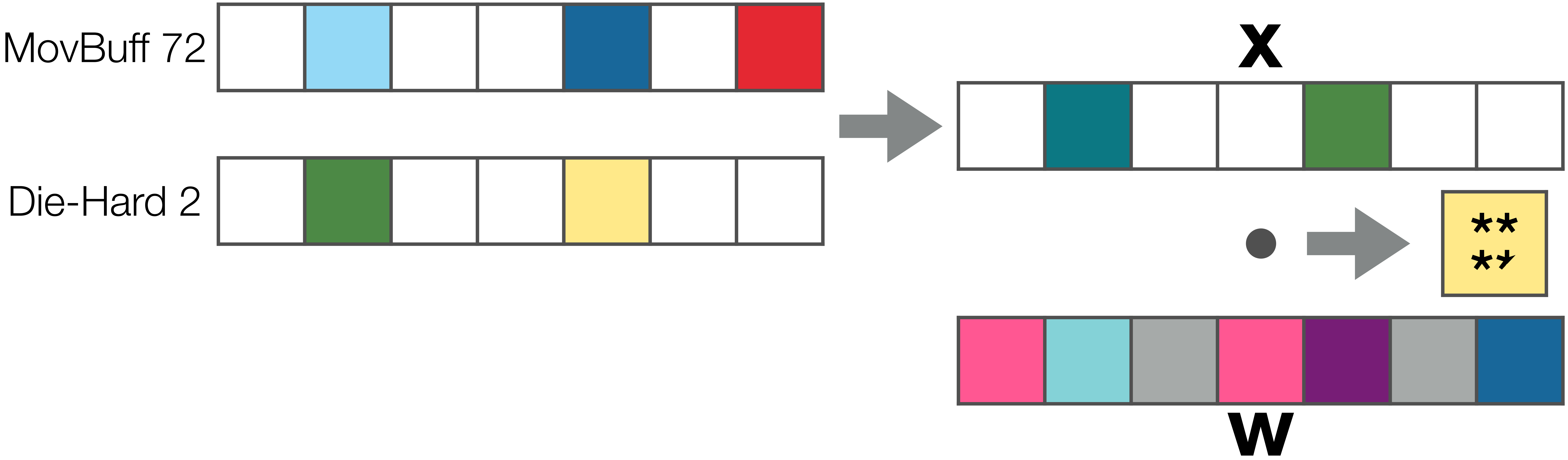

Friends' Favorites
Based on these friends:
Watched by your friends
Daniel Jacobson
John Ciancutti
Mark White
mike Kail


Viewers
Movies
V1
V2
V3
Die Hard
**
***
*
Ex Machina
**
****
?
Magnolia
?
**
**

Linear Regression @ Netflix

- Each viewers is associate with attributes:
  E.g. what genres she/he likes (bag-of-words over genres)
- Each movie is associate with attribute:
  E.g. what genres the movie belongs to (rom-com)
- Learn weight for each pair:
  (Viewer-Attribute-Value , Movie-Attribute-Value)
- Use the star-rating as targets
- For each new movie aggregate its attributes (dark-com)
- Now we can fill-in the gaps for all users

Recommendation Algo



Team Dinosaur Planet

David Lin (Mathematics)
Lester Mackey (Computer Science)
David Weiss (Computer Science)

The Netflix Prize was a $1 million competition to most accurately predict the ratings that people give to the movies they watch. The three of us began working on the Prize in 2006 as undergraduates at Princeton University. In the years since, we helped form the collaborative teams When Gravity and Dinosaurs Unite, Grand Prize Team, and The Ensemble. The contest came to a thrilling conclusion in July of 2009 with The Ensemble placing first on the Quiz Set and second on the Test Set. You can view the results of the competition on the Netflix Prize Leaderboard.

Contact us at <teamdinosaurplanet At gmail DoT com>.

Final Quiz Set Leaderboard

Rank	Team Name	Best Quiz Score	% Improvement	Best Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
Grand Prize - RMSE = 0.8554 - Winning Team: BellKor's Pragmatic Chaos				
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
3	Grand Prize Team	0.8571	9.91	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-10 01:12:31
5	Vandelay Industries !	0.8579	9.83	2009-07-10 00:32:20

Final Test Set Leaderboard

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20

CS Guide Directory Contact

Research People About

s and Events / News

Undergraduates Challenge For
0 NetFlix Prize

Printer Friendly

grads from Princeton's Computer Science Department? When it comes to solving the hardest
eat just about anyone. This year, Lester Mackey '07, David Weiss '07 and David Lin (Math '07)
lix challenge to improve their movie rating system. Despite stiff competition against 2000 other
f whom were professionals with Ph.D.s, the Princeton trio was actually leading the competition 1
petition deadline on October 1st, 2007. At the last moment, they were overtaken by a group of
&T labs and the University of Toronto and wound up finishing second -- an unbelievable result
ing their free time to compete against the best the rest of the world could offer. Read the full
n Paw.

How Analytics Has Given Netflix The Edge Over Hollywood

Enrique Dans

A number of recent articles discuss Hollywood's concern over the recent wave of multimillion dollar Netflix deals with stars like Shonda Rhimes, Ryan Murphy or the Obamas to produce content for the platform. In contrast to the dynamism of Netflix, the traditional movie industry is hamstrung by a business model that depends heavily on sequels, prequels and remakes of popular movies from several decades ago and where success seemingly depends on random or unknown factors.