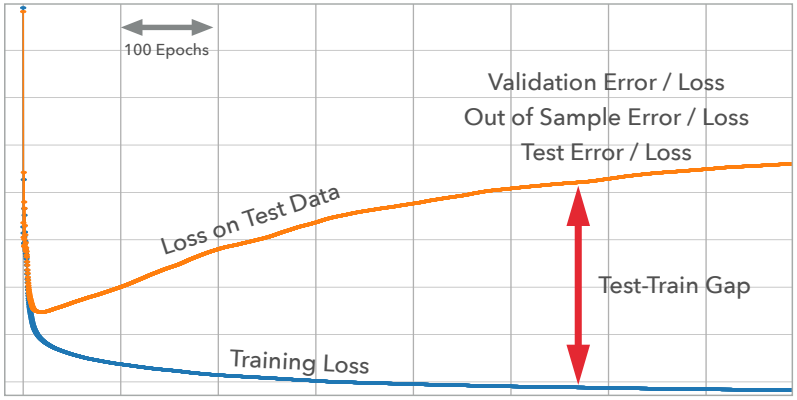
COS324:
Introduction To Machine Learning

Prof. Yoram Singer

PRINCETON UNIVERSITY

Topic: Learning with Regularization

Overfitting



Generalization Analysis: Ingredients

Examples are **I.I.D** samples from an **unknown** distribution $D(\mathbf{x}, y)$

Restrict to $\mathbf{x} \in \mathcal{X} = \{0,1\}^d$ and $y \in \{-1,1\}$

Marginal distribution: $D(\mathbf{x}) = D(\mathbf{x}, -1) + D(\mathbf{x}, +1)$

Conditional distribution: $D(y \mid \mathbf{x}) = \dfrac{D(\mathbf{x}, y)}{D(\mathbf{x})} \in [0,1]$ also denoted $\mathbf{P}(y \mid \mathbf{x})$

Realizable case: $\exists h^\star : \mathcal{X} \to \{-1,1\} \quad s.t. \quad \mathbf{P}\big(y = h^\star(\mathbf{x}) \mid \mathbf{x}\big) = 1$

Unrealizable case: $0 < \mathbf{P}(y \mid \mathbf{x}) < 1$

I.I.D Samples

- I.I.D: Identically Independently Distributed
- Generalization analysis typically assumes $\exists D$:

  unknown distribution $D(\mathbf{x},y)$
- Assume again that $\mathbf{x} \in \{0,1\}^d \quad y \in \{-1,1\}$
- Identically [no dependence on i]:

  $\forall i \in S : \; D((\mathbf{x}_i, y_i) = (\mathbf{a}, b))$ is $D(\mathbf{a}, b)$
- Independence:

  $D\big((\mathbf{x}_i, y_i) = (\mathbf{a}, b) \;\wedge\; D(\mathbf{x}_j, y_j) = (\mathbf{a}', b')\big) = D(\mathbf{a}, b)\, D(\mathbf{a}', b')$

x_0	x_1	y	D
0	0	-1	0.1
0	0	1	0.2
0	1	-1	0
0	1	1	0.3
1	0	-1	0.1
1	0	1	0.2
1	1	-1	0.07
1	1	1	0.03

Unrealizable

x_0	x_1	y	D
0	0	-1	0.1
0	0	1	0.2
0	1	-1	0
0	1	1	0.15
1	0	-1	0.05
1	0	1	0.2
1	1	-1	0.17
1	1	1	0.13

Realizable

x_0	x_1	y	D
0	0	-1	0
0	0	1	0.25
0	1	-1	0
0	1	1	0.2
1	0	-1	0
1	0	1	0.25
1	1	-1	0.3
1	1	1	0

Unrealizable

x_0	x_1	y	D
0	0	-1	0.1
0	0	1	0.2
0	1	-1	0
0	1	1	0.15
1	0	-1	0.05
1	0	1	0.2
1	1	-1	0.17
1	1	1	0.13

Realizable

x_0	x_1	y	D
0	0	-1	0
0	0	1	0.25
0	1	-1	0
0	1	1	0.2
1	0	-1	0
1	0	1	0.25
1	1	-1	0.3
1	1	1	0

Unrealizable

x_0	x_1	y	D
0	0	-1	0.1
0	0	1	0.2
0	1	-1	0
0	1	1	0.15
1	0	-1	0.05
1	0	1	0.2
1	1	-1	0.17
1	1	1	0.13

Realizable

x_0	x_1	y	D
0	0	-1	0
0	0	1	0.25
0	1	-1	0
0	1	1	0.2
1	0	-1	0
1	0	1	0.25
1	1	-1	0.3
1	1	1	0

Unrealizable

x_0	x_1	y	D
0	0	-1	0.1
0	0	1	0.2
0	1	-1	0
0	1	1	0.15
1	0	-1	0.05
1	0	1	0.2
1	1	-1	0.17
1	1	1	0.13

Realizable

x_0	x_1	y	D
0	0	-1	0
0	0	1	0.25
0	1	-1	0
0	1	1	0.2
1	0	-1	0
1	0	1	0.25
1	1	-1	0.3
1	1	1	0

$\equiv$

Realizable

x_0	x_1	$h^\star$	D
0	0	1	0.25
0	1	1	0.2
1	0	1	0.25
1	1	-1	0.3

x_0	x_1	$h^\star$	D
0	0	1	0.25
0	1	1	0.2
1	0	1	0.25
1	1	-1	0.3

x_0	x_1	$h^\star$	D
0	0	1	0.25
0	1	1	0.2
1	0	1	0.25
1	1	-1	0.3

x_0	x_1	y	S
0	0	-1	0
0	0	1	1
0	1	-1	0
0	1	1	1
1	0	-1	0
1	0	1	2
1	1	-1	2
1	1	1	0

YORAM SINGER © 2020 6

Empirical & Population Risk

Empirical risk (or error or loss) of **f** w.r.t. dataset of examples S:

$$\text{err}_S(f) = \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \mathbf{1}\left[f(\mathbf{x}) \neq y\right] \quad \text{// \#errors divided by sample size}$$

Population risk (or error or loss) of **f** w.r.t unknown distribution D:

$$\text{err}_D(f) = \sum_{\mathbf{x}} \sum_{y} D(\mathbf{x}, y)\, \mathbf{1}\left[f(\mathbf{x}) \neq y\right]$$

- Definitions the same for both realizable case & unrealizable case
- Realizable case also amounts to: $\text{err}_D(f) = \sum_{\mathbf{x}} D(\mathbf{x}, y)\, \mathbf{1}\left[f(\mathbf{x}) \neq h^\star(\mathbf{x})\right]$

x_0	x_1	y	D
0	0	-1	0.1
0	0	1	0.2
0	1	-1	0
0	1	1	0.15
1	0	-1	0.05
1	0	1	0.1
1	1	-1	0.27
1	1	1	0.13

f(x)
1
1
1
1
-1
-1
-1
-1

x_0	x_1	y	S
0	0	-1	0
0	0	1	1
0	1	-1	0
0	1	1	1
1	0	-1	0
1	0	1	1
1	1	-1	1
1	1	1	0

$$\mathrm{err}_D(f) = \sum_{x,y} D(x,y)\,\mathbf{1}\left[f(x) \neq y\right] = 0.33$$

x_0	x_1	y	D	f(x)
0	0	-1	0.1	1
0	0	1	0.2	1
0	1	-1	0	1
0	1	1	0.15	1
1	0	-1	0.05	-1
1	0	1	0.1	-1
1	1	-1	0.27	-1
1	1	1	0.13	-1

x_0	x_1	y	S
0	0	-1	0
0	0	1	1
0	1	-1	0
0	1	1	1
1	0	-1	0
1	0	1	1
1	1	-1	1
1	1	1	0

$$\mathrm{err}_D(f) = \sum_{x,y} D(x,y)\,\mathbf{1}\left[f(x) \neq y\right] = 0.33 \qquad \mathrm{err}_S(f) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbf{1}\left[f(x) \neq y\right] = \frac{1}{4}$$

x_0	x_1	y	D	f(x)
0	0	-1	0.1	1
0	0	1	0.2	1
0	1	-1	0	1
0	1	1	0.15	1
1	0	-1	0.05	-1
1	0	1	0.1	-1
1	1	-1	0.27	-1
1	1	1	0.13	-1

x_0	x_1	y	S
0	0	-1	0
0	0	1	1
0	1	-1	0
0	1	1	1
1	0	-1	0
1	0	1	1
1	1	-1	1
1	1	1	0

Finite Set of Predictors

Only **k** predictors — no parameter learning: $f_1, \ldots, f_k$

Finite Set of Predictors

Only **k** predictors — no parameter learning: $f_1, \ldots, f_k$

Best predictor is f_j such that,

$$\mathrm{err}_{\mathbf{D}}[f_j(\mathbf{x})] = \epsilon_j \qquad \forall i \neq j: \ \mathrm{err}_{\mathbf{D}}[f_i(\mathbf{x}) \neq y] \geq \epsilon_j + \epsilon$$

Finite Set of Predictors

Only **k** predictors — no parameter learning: $f_1, \ldots, f_k$

Best predictor is f_j such that,

$$\text{err}_{\mathbf{D}}[f_j(\mathbf{x})] = \epsilon_j \qquad \forall i \neq j : \ \text{err}_{\mathbf{D}}[f_i(\mathbf{x}) \neq y] \geq \epsilon_j + \epsilon$$

Evaluate errors on S: $\text{err}_S(f_r) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[f_r(\mathbf{x}_i) \neq y_i\right]$

Finite Set of Predictors

Only **k** predictors — no parameter learning: $f_1, \ldots, f_k$

Best predictor is f_j such that,

$$\text{err}_{\mathbf{D}}[f_j(\mathbf{x})] = \epsilon_j \qquad \forall i \neq j : \ \text{err}_{\mathbf{D}}[f_i(\mathbf{x}) \neq y] \geq \epsilon_j + \epsilon$$

Evaluate errors on S: $\text{err}_S(f_r) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[f_r(\mathbf{x}_i) \neq y_i\right]$

Choose $\hat{f}$ with smallest empirical risk: $\hat{f} = \arg\min_i \text{err}_S(f_i)$

Generalization

✓ When $\epsilon_j = 0$, if |S| is greater than $\frac{2\log(k)}{\epsilon}$,

then with very high probability $\hat{f} = f_j$

✓ When $\epsilon_j > 0$, if |S| is greater than $\frac{4\log(k)}{\epsilon^2}$

then with very high probability $\hat{f} = f_j$

Overview of Analysis for $\epsilon_j = 0$

Probability that $\epsilon_i = 0$ is at most $(1-\epsilon)^n \leq e^{-\epsilon n}$ [independence of sample]

Overview of Analysis for $\epsilon_j = 0$

Probability that $\epsilon_i = 0$ is at most $(1-\epsilon)^n \leq e^{-\epsilon n}$ [independence of sample]

Probability α that $\exists i \neq j \; s.t. \; \epsilon_i = 0$ is at most $\alpha = (k-1)\, e^{-\epsilon n}$

Overview of Analysis for $\epsilon_j = 0$

Probability that $\epsilon_i = 0$ is at most $(1-\epsilon)^n \leq e^{-\epsilon n}$ [independence of sample]

Probability α that $\exists i \neq j \; s.t. \; \epsilon_i = 0$ is at most $\alpha = (k-1)\, e^{-\epsilon n}$

If $\alpha \leq \frac{1}{k}$ it is unlikely we do not find correct predictor: $(k-1)\, e^{-\epsilon n} \geq \frac{1}{k}$

Overview of Analysis for $\epsilon_j = 0$

Probability that $\epsilon_i = 0$ is at most $(1-\epsilon)^n \leq e^{-\epsilon n}$ [independence of sample]

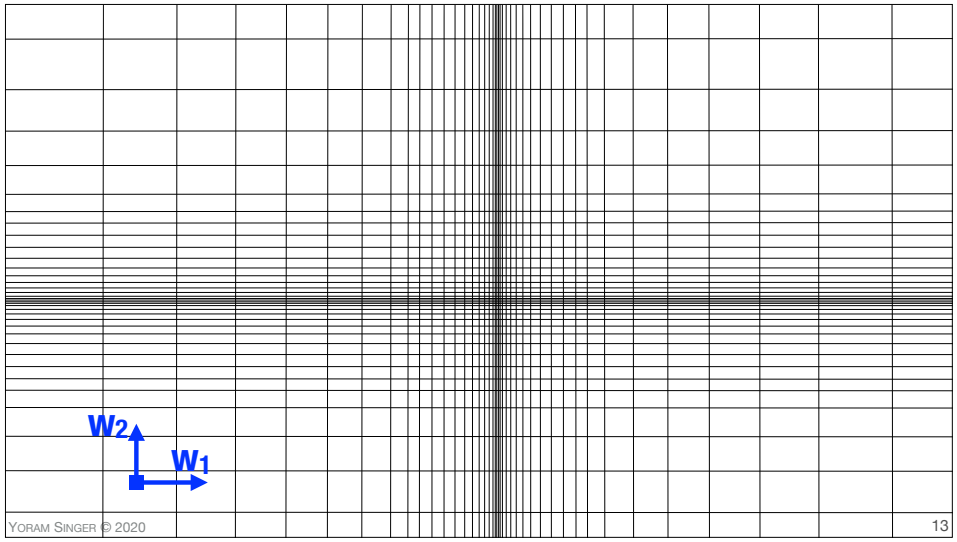
Probability α that $\exists i \neq j \; s.t. \; \epsilon_i = 0$ is at most $\alpha = (k-1)\, e^{-\epsilon n}$

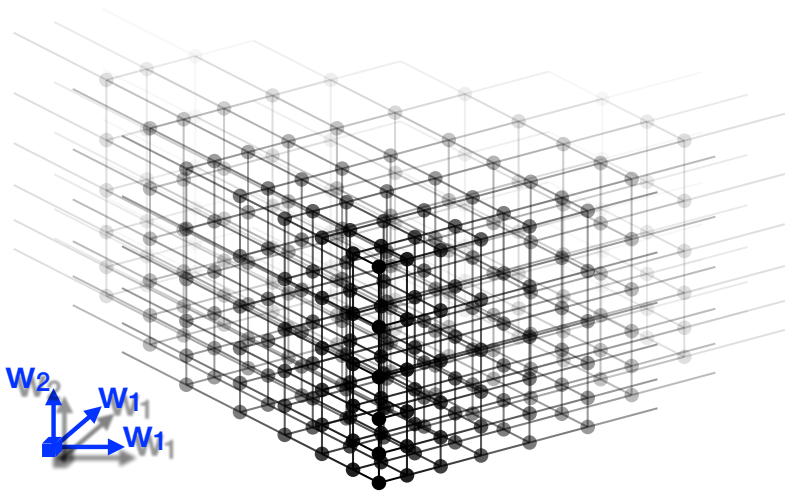
If $\alpha \leq \frac{1}{k}$ it is unlikely we do not find correct predictor: $(k-1)\, e^{-\epsilon n} \geq \frac{1}{k}$

This means that we need $O\left(\frac{\log(k)}{\epsilon}\right)$ samples

Finite Precision Reals (Floating Point)

Also : Inf $\equiv \infty$ and NaN





Brute Force Enumeration

Given say a neural net structure with **d** parameters

Brute Force Enumeration

Given say a neural net structure with **d** parameters

Enumerate all possible floating point values for **d** parameters

Brute Force Enumeration

Given say a neural net structure with **d** parameters

Enumerate all possible floating point values for **d** parameters

Each parameter has less than $\mathbf{2^{32}}$ different values

Brute Force Enumeration

Given say a neural net structure with **d** parameters

Enumerate all possible floating point values for **d** parameters

Each parameter has less than $\mathbf{2^{32}}$ different values

We have "only" 2^{32d} different neural nets $\Rightarrow f_1, \ldots, f_{2^{32d}}$

Brute Force Enumeration

Given say a neural net structure with **d** parameters

Enumerate all possible floating point values for **d** parameters

Each parameter has less than $\mathbf{2^{32}}$ different values

We have "only" 2^{32d} different neural nets $\Rightarrow f_1, \ldots, f_{2^{32d}}$

Only $\tilde{O}(d)$ examples needed to find w.h.p NN w/ the best **generalization**

Brute Force Enumeration

Given say a neural net structure with **d** parameters

Enumerate all possible floating point values for **d** parameters

Each parameter has less than $\mathbf{2^{32}}$ different values

We have "only" 2^{32d} different neural nets $\Rightarrow f_1, \ldots, f_{2^{32d}}$

Only $\tilde{O}(d)$ examples needed to find w.h.p NN w/ the best **generalization**

- We need $O(\exp(d))$ time

Brute Force Enumeration

Given say a neural net structure with **d** parameters

Enumerate all possible floating point values for **d** parameters

Each parameter has less than 2^{32} different values

We have "only" 2^{32d} different neural nets $\Rightarrow f_1, \ldots, f_{2^{32d}}$

Only $\tilde{O}(d)$ examples needed to find w.h.p NN w/ the best **generalization**

- We need $O(\exp(d))$ time
- $\tilde{O}(d)$ hides really big constants$(\gg 32)$ because of ϵ

Brute Force Enumeration

Given say a neural net structure with **d** parameters

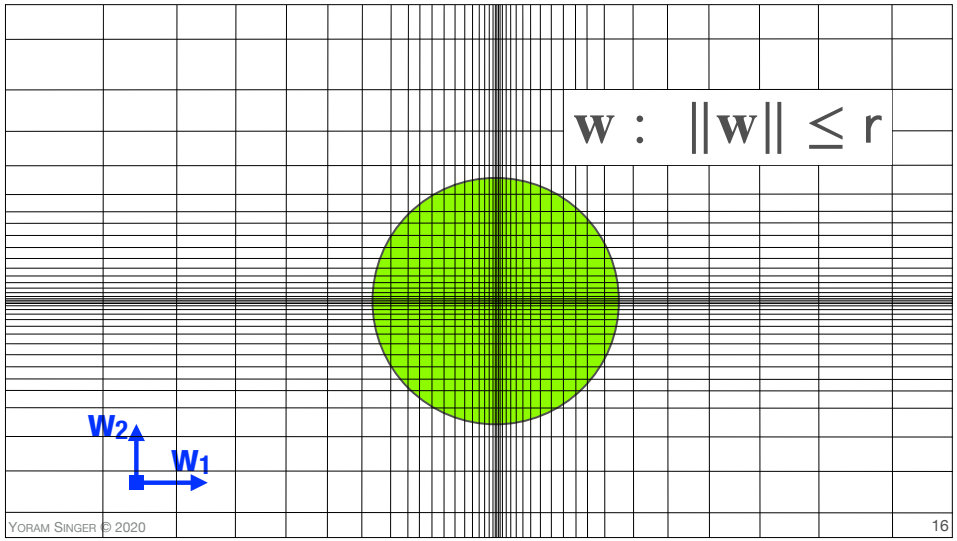
Enumerate all possible floating point values for **d** parameters

Each parameter has less than 2^{32} different values

We have "only" 2^{32d} different neural nets $\Rightarrow f_1, \ldots, f_{2^{32d}}$

Only $\tilde{O}(d)$ examples needed to find w.h.p NN w/ the best **generalization**

- We need $O(\exp(d))$ time
- $\tilde{O}(d)$ hides really big constants$(\gg 32)$ because of ϵ
- For any reasonable amount of examples we are doomed to overfit



Projection Onto a Ball

Limits space of admissible parameters

Helps in preventing overfitting

Simple & seamless to implement w/ SGD

Projection Onto a Ball

Limits space of admissible parameters

Helps in preventing overfitting

Simple & seamless to implement w/ SGD

However, semantics of radius is not clear

Projection Onto a Ball

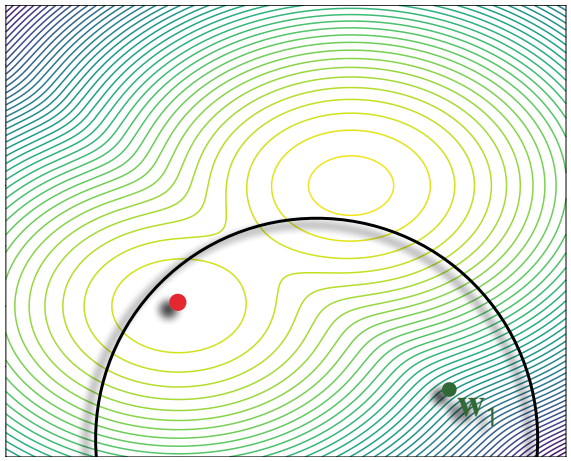
Limits space of admissible parameters

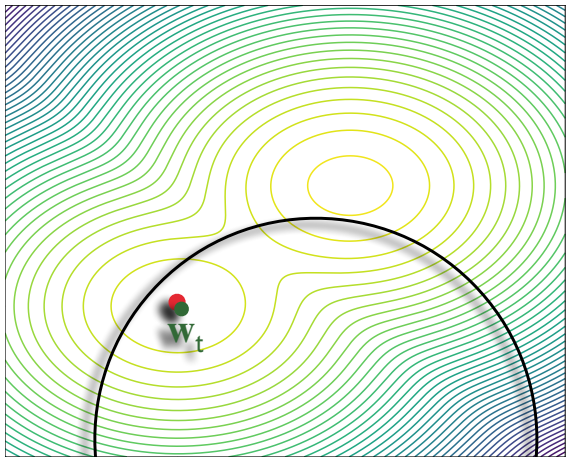
Helps in preventing overfitting

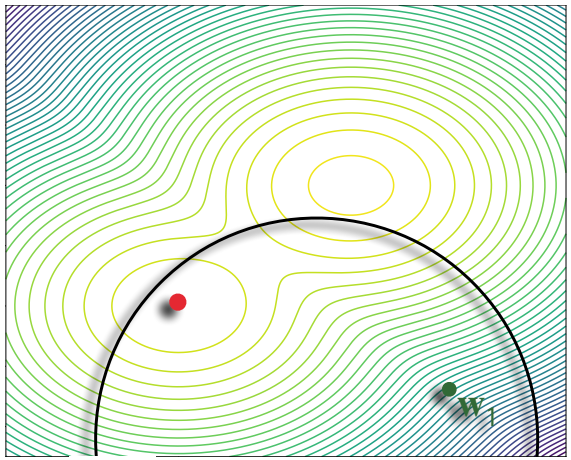
Simple & seamless to implement w/ SGD

However, semantics of radius is not clear

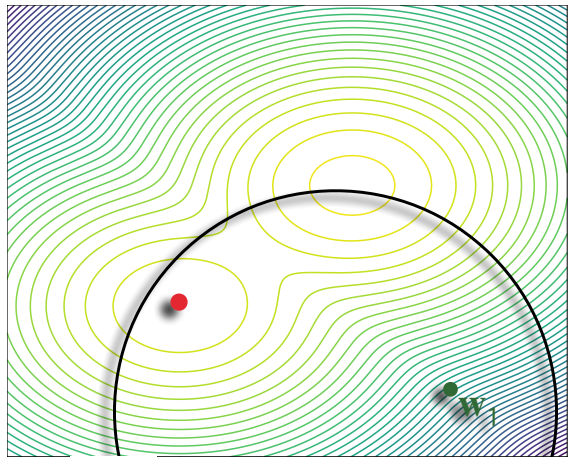
Often ill-defined in non-linear models (e.g NN)



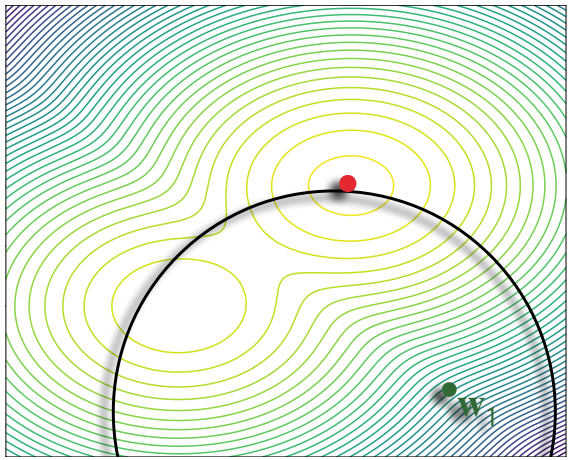




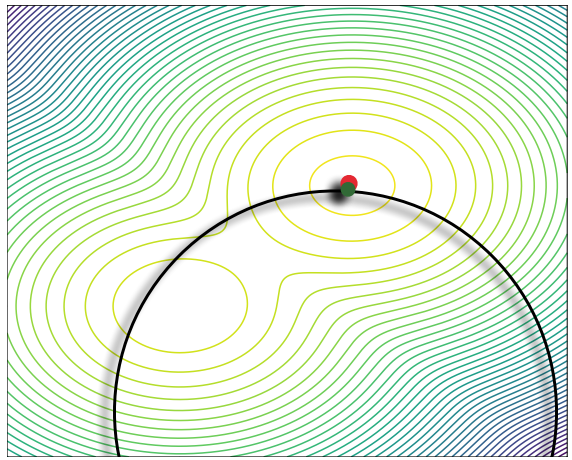
Constrained Optimum



Constrained Optimum



Constrained Optimum



Constrained Optimum

Regularization

Instead of constrained problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\| \le r$$

Solve penalized problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|^2 = \mathscr{L}(\mathbf{w}) + \lambda \sum_j w_j^2$$

Regularization

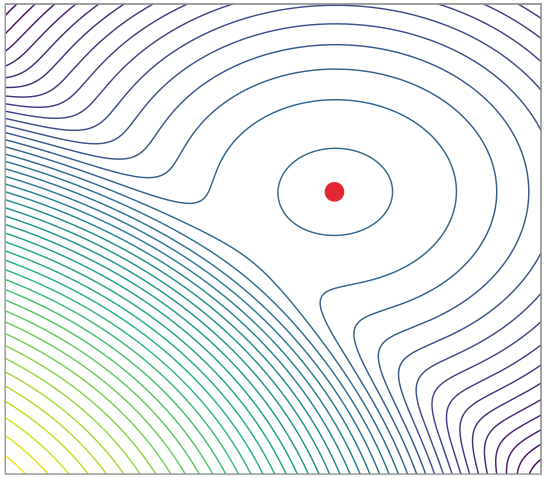
Instead of constrained problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\| \le r$$

Solve penalized problem:

regularization value

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|^2 = \mathscr{L}(\mathbf{w}) + \lambda \sum_j w_j^2$$



Regularization ⇒ ∃Domain Constraints

Recall that: $\mathscr{L}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left((\mathbf{x}_i, y_i), \mathbf{w}\right) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell_i(\mathbf{w})$

Regularization $\Rightarrow$ $\exists$Domain Constraints

Recall that: $\mathscr{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((\mathbf{x}_i, y_i), \mathbf{w}\right) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w})$

By construction: $\ell_i(\mathbf{0}) = 1 \;\Rightarrow\; \mathscr{L}(\mathbf{0}) = 1$ and $\ell_i(\mathbf{w}) \geq 0 \;\Rightarrow\; \mathscr{L}(\mathbf{w}) \geq 0$

Regularization $\Rightarrow$ $\exists$Domain Constraints

Recall that: $\mathscr{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((\mathbf{x}_i, y_i), \mathbf{w}\right) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w})$

By construction: $\ell_i(\mathbf{0}) = 1 \;\Rightarrow\; \mathscr{L}(\mathbf{0}) = 1$ and $\ell_i(\mathbf{w}) \geq 0 \;\Rightarrow\; \mathscr{L}(\mathbf{w}) \geq 0$

Since we look for a (local) minimum:

$$1 = \mathscr{L}(\mathbf{0}) + \lambda\|\mathbf{0}\|^2 \geq \mathscr{L}(\mathbf{w}^\star) + \lambda\|\mathbf{w}^\star\|^2 \geq \lambda\|\mathbf{w}^\star\|^2$$

Regularization $\Rightarrow$ $\exists$Domain Constraints

Recall that: $\mathscr{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((\mathbf{x}_i, y_i), \mathbf{w}\right) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w})$

By construction: $\ell_i(\mathbf{0}) = 1 \;\Rightarrow\; \mathscr{L}(\mathbf{0}) = 1$ and $\ell_i(\mathbf{w}) \geq 0 \;\Rightarrow\; \mathscr{L}(\mathbf{w}) \geq 0$

Since we look for a (local) minimum:

$$1 = \mathscr{L}(\mathbf{0}) + \lambda\|\mathbf{0}\|^2 \geq \mathscr{L}(\mathbf{w}^\star) + \lambda\|\mathbf{w}^\star\|^2 \geq \lambda\|\mathbf{w}^\star\|^2$$

Therefore $\mathbf{w}^\star$ implicitly satisfies the ball constraint $\|\mathbf{w}^\star\| \leq \frac{1}{\sqrt{\lambda}}$

Regularization $\Rightarrow$ $\exists$Domain Constraints

Recall that: $\mathscr{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((\mathbf{x}_i, y_i), \mathbf{w}\right) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w})$

By construction: $\ell_i(\mathbf{0}) = 1 \;\Rightarrow\; \mathscr{L}(\mathbf{0}) = 1$ and $\ell_i(\mathbf{w}) \geq 0 \;\Rightarrow\; \mathscr{L}(\mathbf{w}) \geq 0$

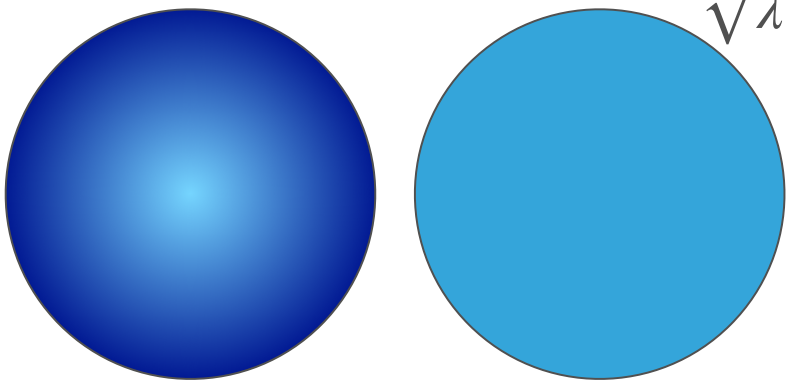
Since we look for a (local) minimum:

$$1 = \mathscr{L}(\mathbf{0}) + \lambda\|\mathbf{0}\|^2 \geq \mathscr{L}(\mathbf{w}^\star) + \lambda\|\mathbf{w}^\star\|^2 \geq \lambda\|\mathbf{w}^\star\|^2$$

Therefore $\mathbf{w}^\star$ implicitly satisfies the ball constraint $\|\mathbf{w}^\star\| \leq \frac{1}{\sqrt{\lambda}}$

However, the converse (domain $\Rightarrow$ $\exists$regularization) does not hold

$$\ldots + \lambda \|\mathbf{w}\|^2 \qquad \|\mathbf{w}\| \leq \frac{1}{\sqrt{\lambda}}$$



SGD w/ 2-norm Regularization

$$\frac{\partial}{\partial w_j} \|\mathbf{w}\|^2 = \frac{\partial}{\partial w_j} \sum_i w_i^2 = \frac{\partial w_j^2}{\partial w_j}$$

SGD w/ 2-norm Regularization

Note that $\nabla_{\mathbf{w}} \|\mathbf{w}\|^2 = \left(\frac{\partial w_1^2}{\partial w_1}, \ \ldots \ , \frac{\partial w_d^2}{\partial w_d} \right) = 2 \left(w_1, \ \ldots \ , w_d \right) = 2\mathbf{w}$

SGD w/ 2-norm Regularization

Note that $\nabla_{\mathbf{w}} \|\mathbf{w}\|^2 = \left(\frac{\partial w_1^2}{\partial w_1}, \ \ldots \ , \frac{\partial w_d^2}{\partial w_d} \right) = 2 \left(w_1, \ \ldots \ , w_d \right) = 2\mathbf{w}$

Add $\lambda \|\mathbf{w}\|^2$ to the objective $\nabla_{\mathbf{w}} \left[\mathscr{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|^2 \right] = \nabla \mathscr{L}(\mathbf{w}) + 2\lambda \mathbf{w}$

SGD w/ 2-norm Regularization

Note that $\nabla_{\mathbf{w}} \|\mathbf{w}\|^2 = \left(\frac{\partial w_1^2}{\partial w_1}, \ldots, \frac{\partial w_d^2}{\partial w_d}\right) = 2(w_1, \ldots, w_d) = 2\mathbf{w}$

Add $\lambda\|\mathbf{w}\|^2$ to the objective $\nabla_{\mathbf{w}} \left[\mathscr{L}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2\right] = \nabla \mathscr{L}(\mathbf{w}) + 2\lambda\mathbf{w}$

In stochastic case we get $\frac{1}{|S|} \sum_{i \in S} \nabla \ell_i(\mathbf{w}_t) + 2\lambda\mathbf{w}_t$

SGD w/ 2-norm Regularization

Note that $\nabla_{\mathbf{w}} \|\mathbf{w}\|^2 = \left(\frac{\partial w_1^2}{\partial w_1}, \ldots, \frac{\partial w_d^2}{\partial w_d}\right) = 2(w_1, \ldots, w_d) = 2\mathbf{w}$

Add $\lambda\|\mathbf{w}\|^2$ to the objective $\nabla_{\mathbf{w}} \left[\mathscr{L}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2\right] = \nabla \mathscr{L}(\mathbf{w}) + 2\lambda\mathbf{w}$

In stochastic case we get $\frac{1}{|S|} \sum_{i \in S} \nabla \ell_i(\mathbf{w}_t) + 2\lambda\mathbf{w}_t$

$\mathbf{g}_t$

SGD w/ 2-norm Regularization

Note that $\nabla_{\mathbf{w}} \|\mathbf{w}\|^2 = \left(\frac{\partial w_1^2}{\partial w_1}, \ldots, \frac{\partial w_d^2}{\partial w_d}\right) = 2(w_1, \ldots, w_d) = 2\mathbf{w}$

Add $\lambda\|\mathbf{w}\|^2$ to the objective $\nabla_{\mathbf{w}} \left[\mathscr{L}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2\right] = \nabla \mathscr{L}(\mathbf{w}) + 2\lambda\mathbf{w}$

In stochastic case we get $\frac{1}{|S|} \sum_{i \in S} \nabla \ell_i(\mathbf{w}_t) + 2\lambda\mathbf{w}_t$

SGD update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t + 2\lambda\mathbf{w}_t) = (1 - 2\lambda\eta_t)\mathbf{w}_t - \eta_t\mathbf{g}_t$

SGD w/ 2-norm Regularization

Note that $\nabla_{\mathbf{w}} \|\mathbf{w}\|^2 = \left(\frac{\partial w_1^2}{\partial w_1}, \ldots, \frac{\partial w_d^2}{\partial w_d}\right) = 2(w_1, \ldots, w_d) = 2\mathbf{w}$

Add $\lambda\|\mathbf{w}\|^2$ to the objective $\nabla_{\mathbf{w}} \left[\mathscr{L}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2\right] = \nabla \mathscr{L}(\mathbf{w}) + 2\lambda\mathbf{w}$

In stochastic case we get $\frac{1}{|S|} \sum_{i \in S} \nabla \ell_i(\mathbf{w}_t) + 2\lambda\mathbf{w}_t$

weight decay

SGD update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t + 2\lambda\mathbf{w}_t) = (1 - 2\lambda\eta_t)\mathbf{w}_t - \eta_t\mathbf{g}_t$

Suppose $\ell\big(y(\mathbf{w}\cdot\mathbf{x})\big) = [1 - y(\mathbf{w}\cdot\mathbf{x})]_+$

$$\mathscr{L}\big([1,0]\big) = \frac{1}{4}(1+0+0+2) = \frac{3}{4}$$

$$\mathscr{L}\big([0,1]\big) = \frac{1}{4}(2+0+0+1) = \frac{3}{4}$$

$$\mathscr{L}\big([0.5,0.5]\big) = \frac{1}{4}(1.5+0+0+1.5) = \frac{3}{4}$$

$$\mathscr{L}\big([1,0]\big) = \mathscr{L}\big([0,1]\big) = \mathscr{L}\big([0.5,0.5]\big) = \frac{3}{4}$$

	x[0]	x[1]	y
Ex. 1	0	1	-1
Ex. 2	1	1	1
Ex. 3	-1	-1	-1
Ex. 4	-1	0	1

... no preference ...

Regularized version: $\mathscr{L}(\mathbf{w}) + 10^{-6}\,\|\mathbf{w}\|^2$

$$\mathscr{L}\big([0.5,0.5]\big) + 10^{-6}\,[0.5^2 + 0.5^2] = \frac{3}{4} + \frac{1}{2}10^{-6}$$

$$\mathscr{L}\big([1,0]\big) + 10^{-6}\,[1^2 + 0^2] = \frac{3}{4} + 10^{-6}$$

$$\mathscr{L}\big([0,1]\big) + 10^{-6}\,[0^2 + 1^2] = \frac{3}{4} + 10^{-6}$$

$\|\ \|_2$ regularization is Inclusive

SGD w/ 1-norm Regularization

Recall $\|\mathbf{w}\|_1 = \sum_j |\,\mathbf{w}_j\,|$ and derivative at $\mathbf{w}_j = 0$ is not uniquely defined

SGD w/ 1-norm Regularization

Recall $\|\mathbf{w}\|_1 = \sum_j |\,\mathbf{w}_j\,|$ and derivative at $\mathbf{w}_j = 0$ is not uniquely defined

Start with a simple single variable problem: $\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) + \lambda\mathbf{w}$ s.t $\mathbf{w} \geq 0$

SGD w/ 1-norm Regularization

Recall $\|\mathbf{w}\|_1 = \sum_j |w_j|$ and derivative at $w_j = 0$ is not uniquely defined

Start with a simple single variable problem: $\min_w \mathscr{L}(w) + \lambda w$ s.t $w \geq 0$

Let $g_t = \frac{d\,\mathscr{L}(w)}{dw}\Big|_{w=w_t}$ For $w > 0$ derivative well defined $\frac{d}{dw}\lambda w = \lambda$

SGD w/ 1-norm Regularization

Recall $\|\mathbf{w}\|_1 = \sum_j |w_j|$ and derivative at $w_j = 0$ is not uniquely defined

Start with a simple single variable problem: $\min_w \mathscr{L}(w) + \lambda w$ s.t $w \geq 0$

Let $g_t = \frac{d\,\mathscr{L}(w)}{dw}\Big|_{w=w_t}$ For $w > 0$ derivative well defined $\frac{d}{dw}\lambda w = \lambda$

If w_t and w_{t+1} are positive update is : $w_{t+1} \leftarrow w_t - \eta_t(g_t + \lambda)$

SGD w/ 1-norm Regularization

Recall $\|\mathbf{w}\|_1 = \sum_j |w_j|$ and derivative at $w_j = 0$ is not uniquely defined

Start with a simple single variable problem: $\min_w \mathscr{L}(w) + \lambda w$ s.t $w \geq 0$

Let $g_t = \frac{d\,\mathscr{L}(w)}{dw}\Big|_{w=w_t}$ For $w > 0$ derivative well defined $\frac{d}{dw}\lambda w = \lambda$

If w_t and w_{t+1} are positive update is : $w_{t+1} \leftarrow w_t - \eta_t(g_t + \lambda)$

However if $w_t - \eta_t(g_t + \lambda) \leq 0 \;\Leftrightarrow\; w_t - \eta_t g_t \leq \eta_t\lambda$ we must stop at 0

SGD w/ 1-norm Regularization

Recall $\|\mathbf{w}\|_1 = \sum_j |w_j|$ and derivative at $w_j = 0$ is not uniquely defined

Start with a simple single variable problem: $\min_w \mathscr{L}(w) + \lambda w$ s.t $w \geq 0$

Let $g_t = \frac{d\,\mathscr{L}(w)}{dw}\Big|_{w=w_t}$ For $w > 0$ derivative well defined $\frac{d}{dw}\lambda w = \lambda$

If w_t and w_{t+1} are positive update is : $w_{t+1} \leftarrow w_t - \eta_t(g_t + \lambda)$

However if $w_t - \eta_t(g_t + \lambda) \leq 0 \;\Leftrightarrow\; w_t - \eta_t g_t \leq \eta_t\lambda$ we must stop at 0

$$w_{t+1} \leftarrow \max\left\{w_t - \eta_t g_t - \eta_t\lambda,\, 0\right\}$$

SGD w/ 1-norm Regularization

Similarly constraining $\mathbf{w} \leq 0$ while penalizing for $\lambda |\mathbf{w}|$ yields the problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) - \lambda \mathbf{w} \;\text{ s.t }\; \mathbf{w} \leq 0$$

SGD w/ 1-norm Regularization

Similarly constraining $\mathbf{w} \leq 0$ while penalizing for $\lambda |\mathbf{w}|$ yields the problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) - \lambda \mathbf{w} \;\text{ s.t }\; \mathbf{w} \leq 0$$

Assuming $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ are *negative* update is : $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t - \lambda)$

SGD w/ 1-norm Regularization

Similarly constraining $\mathbf{w} \leq 0$ while penalizing for $\lambda |\mathbf{w}|$ yields the problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) - \lambda \mathbf{w} \;\text{ s.t }\; \mathbf{w} \leq 0$$

Assuming $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ are *negative* update is : $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t - \lambda)$

However if $\mathbf{w}_t - \eta_t(\mathbf{g}_t - \lambda) \geq 0 \;\Leftrightarrow\; \mathbf{w}_t - \eta_t \mathbf{g}_t \geq -\eta_t \lambda$ we must stop at 0

SGD w/ 1-norm Regularization

Similarly constraining $\mathbf{w} \leq 0$ while penalizing for $\lambda |\mathbf{w}|$ yields the problem:

$$\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) - \lambda \mathbf{w} \;\text{ s.t }\; \mathbf{w} \leq 0$$

Assuming $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ are *negative* update is : $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t - \lambda)$

However if $\mathbf{w}_t - \eta_t(\mathbf{g}_t - \lambda) \geq 0 \;\Leftrightarrow\; \mathbf{w}_t - \eta_t \mathbf{g}_t \geq -\eta_t \lambda$ we must stop at 0

SGD w/ 1-norm Regularization

Similarly constraining $w \leq 0$ while penalizing for $\lambda |w|$ yields the problem:

$$\min_w \mathscr{L}(w) - \lambda w \text{ s.t } w \leq 0$$

Assuming w_t and w_{t+1} are *negative* update is : $w_{t+1} \leftarrow w_t - \eta_t(g_t - \lambda)$

However if $w_t - \eta_t(g_t - \lambda) \geq 0 \Leftrightarrow w_t - \eta_t g_t \geq -\eta_t\lambda$ we must stop at 0

$$w_{t+1} \leftarrow \min\left\{w_t - \eta_t g_t + \eta_t\lambda,\, 0\right\}$$

SGD w/ 1-norm Regularization

Combining the two problems:

$$w^+ = \arg\min_w \mathscr{L}(w) + \lambda w \text{ s.t } w \geq 0 \qquad w^- = \arg\min_w \mathscr{L}(w) - \lambda w \text{ s.t } w \leq 0$$

SGD w/ 1-norm Regularization

Combining the two problems:

$$w^+ = \arg\min_w \mathscr{L}(w) + \lambda w \text{ s.t } w \geq 0 \qquad w^- = \arg\min_w \mathscr{L}(w) - \lambda w \text{ s.t } w \leq 0$$

- $w^+ > 0$ & $w^- = 0 \Rightarrow w^\star > 0$

SGD w/ 1-norm Regularization

Combining the two problems:

$$w^+ = \arg\min_w \mathscr{L}(w) + \lambda w \text{ s.t } w \geq 0 \qquad w^- = \arg\min_w \mathscr{L}(w) - \lambda w \text{ s.t } w \leq 0$$

- $w^+ > 0$ & $w^- = 0 \Rightarrow w^\star > 0$
- $w^+ = 0$ & $w^- < 0 \Rightarrow w^\star < 0$

SGD w/ 1-norm Regularization

Combining the two problems:

$$w^+ = \arg\min_w \mathscr{L}(w) + \lambda w \text{ s.t } w \geq 0 \qquad w^- = \arg\min_w \mathscr{L}(w) - \lambda w \text{ s.t } w \leq 0$$

- $w^+ > 0$ & $w^- = 0 \Rightarrow w^\star > 0$
- $w^+ = 0$ & $w^- < 0 \Rightarrow w^\star < 0$
- $w^+ = 0$ & $w^- = 0 \Rightarrow w^\star = 0$

SGD w/ 1-norm Regularization

Construct $\min_w \mathscr{L}(w) + \lambda |w|$ by combining the two cases

SGD w/ 1-norm Regularization

Construct $\min_w \mathscr{L}(w) + \lambda |w|$ by combining the two cases

If $-\eta_t\lambda \leq w_t - \eta_t g_t \leq \eta_t\lambda \Leftrightarrow |w_t - \eta_t g_t| \leq \eta_t\lambda$ then

SGD w/ 1-norm Regularization

Construct $\min_w \mathscr{L}(w) + \lambda |w|$ by combining the two cases

If $-\eta_t\lambda \leq w_t - \eta_t g_t \leq \eta_t\lambda \Leftrightarrow |w_t - \eta_t g_t| \leq \eta_t\lambda$ then

$w_{t+1} = 0$

SGD w/ 1-norm Regularization

Construct $\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) + \lambda |\mathbf{w}|$ by combining the two cases

If $-\eta_t\lambda \le \mathbf{w}_t - \eta_t\mathbf{g}_t \le \eta_t\lambda \;\Leftrightarrow\; |\mathbf{w}_t - \eta_t\mathbf{g}_t| \le \eta_t\lambda$ then

$\mathbf{w}_{t+1} = 0$

If $\mathbf{w}_t - \eta_t\mathbf{g}_t > \eta_t\lambda$ then

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t + \lambda)$

If $\mathbf{w}_t - \eta_t\mathbf{g}_t < -\eta_t\lambda$ then

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t(\mathbf{g}_t - \lambda)$

SGD w/ 1-norm Regularization

To update $\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$ we perform

i. $w_j^{t+1/2} \leftarrow w_j^t - \eta_t g_j^t$

ii. $w_j^{t+1} \leftarrow \left(\text{sign}(w_j^{t+1/2})\right) \left[\, |w_j^{t+1/2}| - \eta_t \lambda \right]_+$

SGD w/ 1-norm Regularization

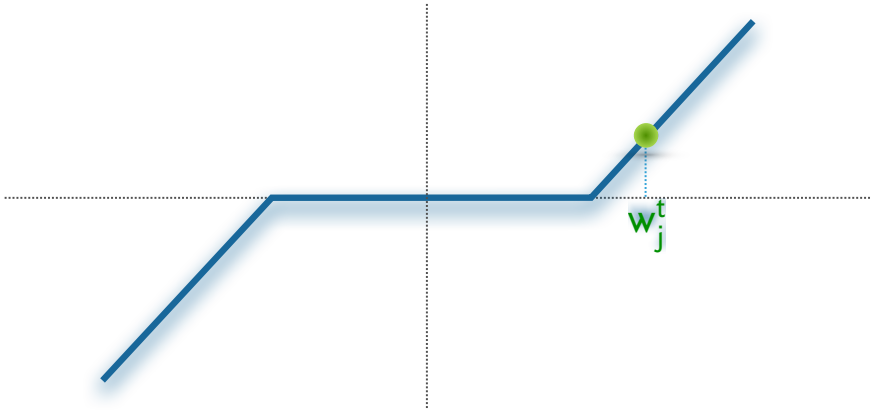
To update $\min_{\mathbf{w}} \mathscr{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$ we perform

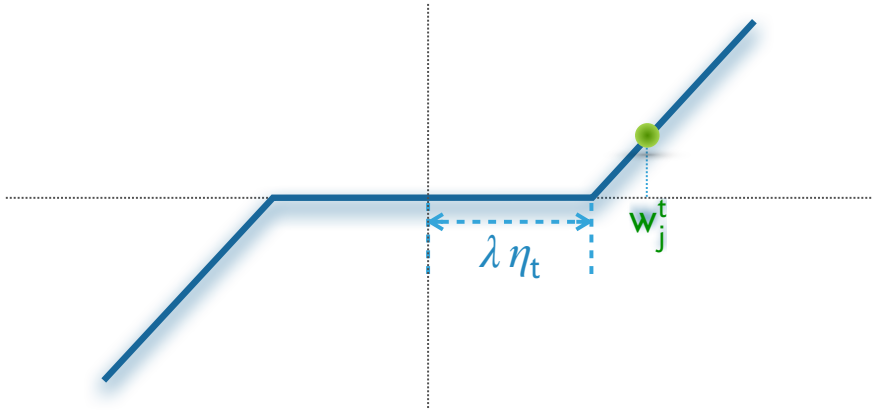
i. $w_j^{t+1/2} \leftarrow w_j^t - \eta_t g_j^t$

ii. $w_j^{t+1} \leftarrow \left(\text{sign}(w_j^{t+1/2})\right) \left[\, |w_j^{t+1/2}| - \eta_t \lambda \right]_+$

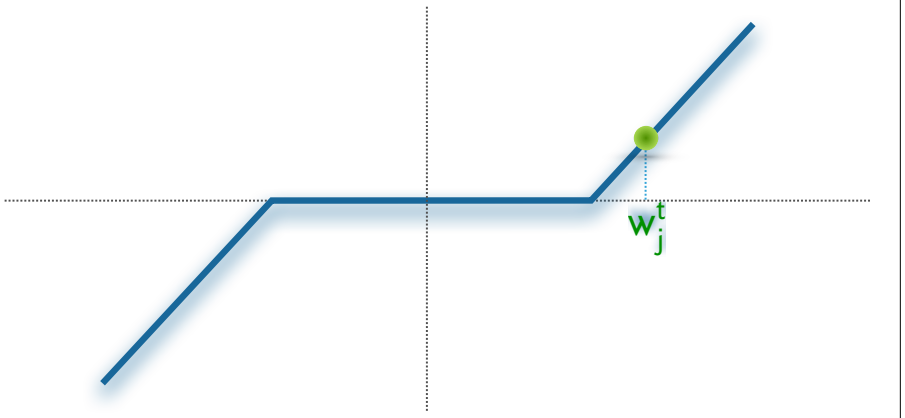
YORAM SINGER © 2020 31

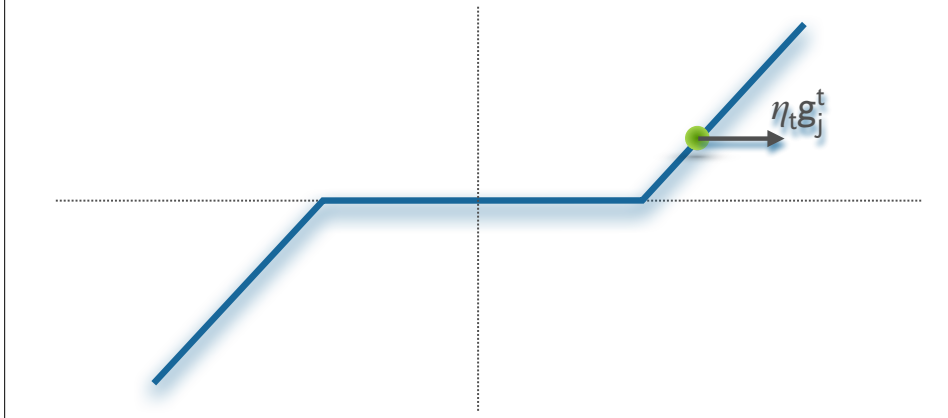
i. $w_j^{t+1/2} \leftarrow w_j^t - \eta_t g_j^t$ ii. $w_j^{t+1} \leftarrow \left(\text{sign}(w_j^{t+1/2})\right) \left[\, |w_j^{t+1/2}| - \eta_t \lambda \right]_+$

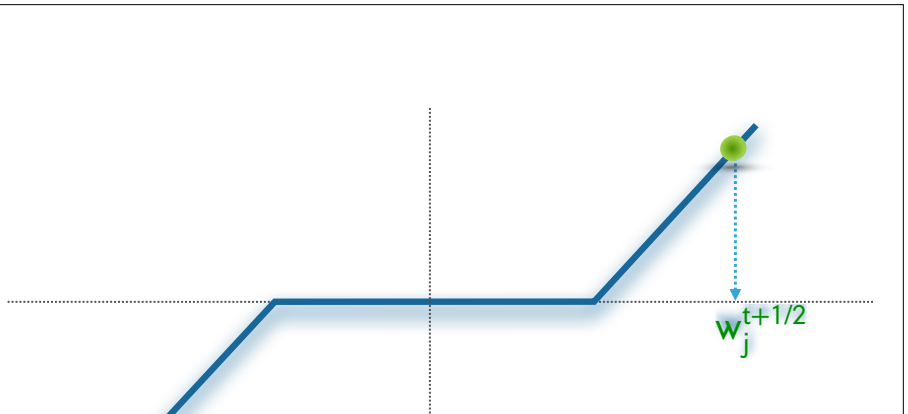


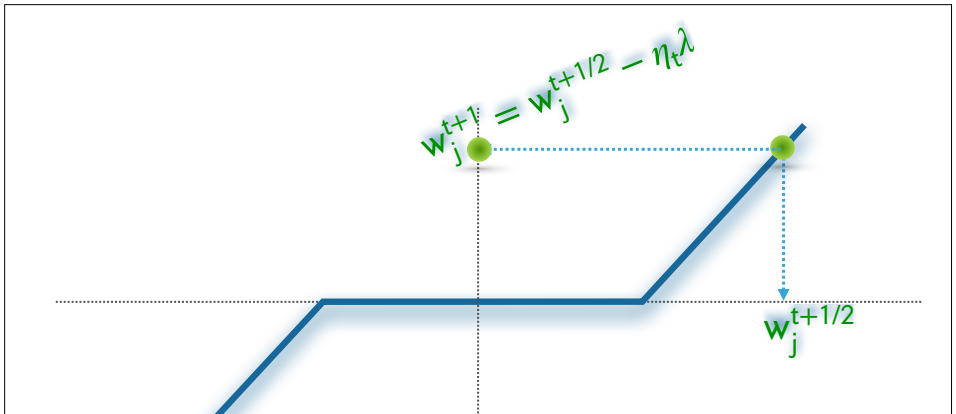


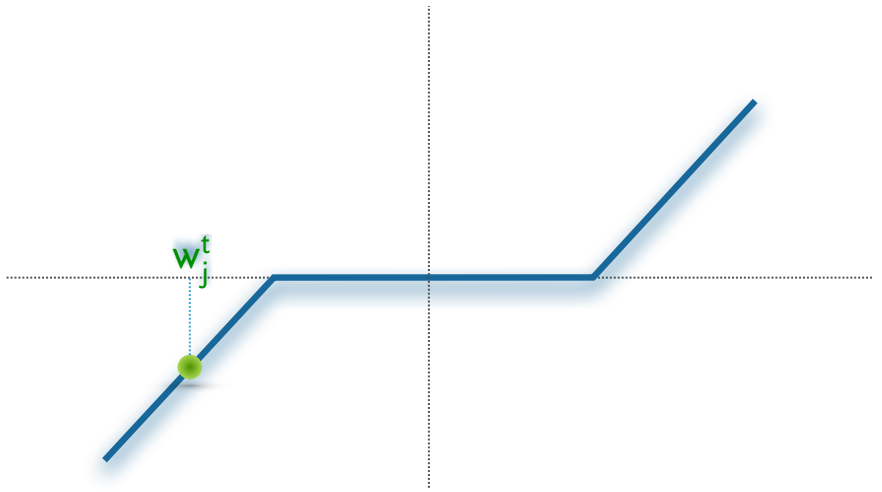
YORAM SINGER © 2020 32

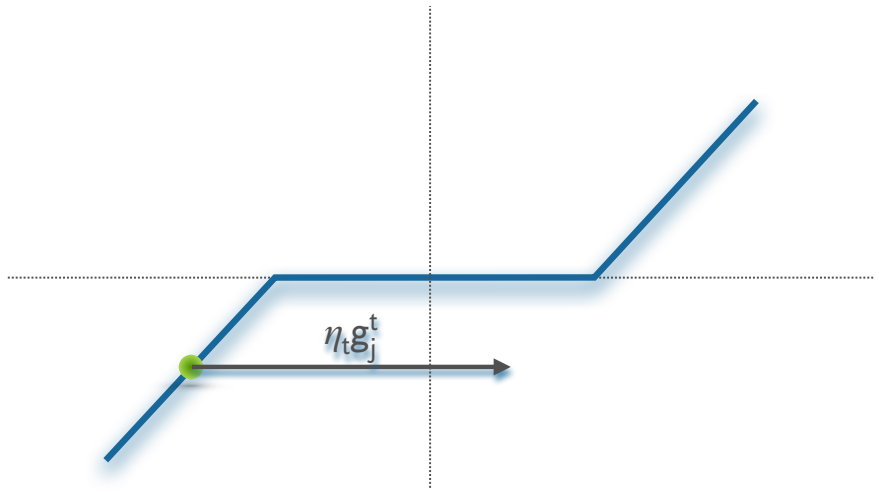

w_j^t

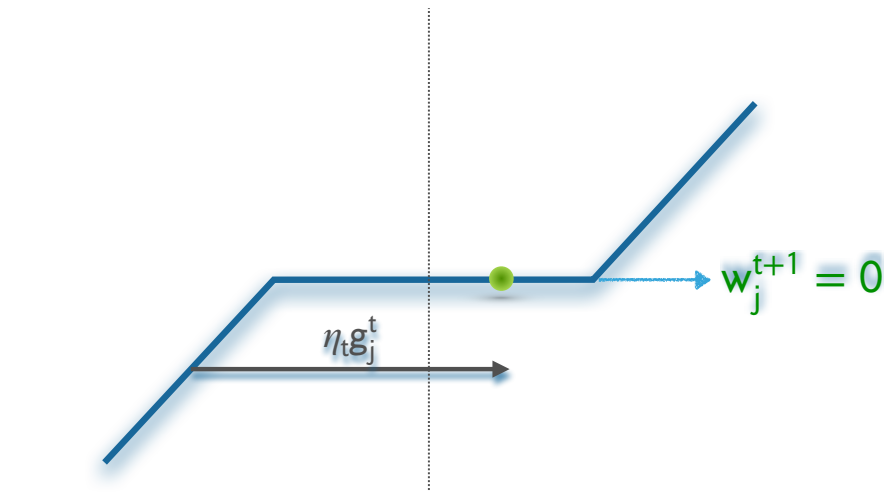

$\eta_t g_j^t$


$w_j^{t+1/2}$


$w_j^{t+1} = w_j^{t+1/2} - \eta_t \lambda$
$w_j^{t+1/2}$







Suppose $\ell\left(y(\mathbf{w}\cdot\mathbf{x})\right) = [1 - y(\mathbf{w}\cdot\mathbf{x})]_+$

$$\mathscr{L}(1,0) = \frac{1}{4}(1+0+0+2) = \frac{3}{4}$$

$$\mathscr{L}(0,1) = \frac{1}{4}(2.08+0+0+1) = \frac{3}{4} + 0.02$$

$$\mathscr{L}(0.5,0.5) = \frac{1}{4}(1.54+0+0+1.5) = \frac{3}{4} + 0.01$$

$$\mathscr{L}(0,1) > \mathscr{L}(0.5,0.5) > \mathscr{L}(1,0)$$

x[0]	x[1]	y
0	1.08	-1
1	1	1
-1	-1	-1
-1	0	1

Regularized objective: $Q(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \|\mathbf{w}\|_1 - \frac{3}{4}$

$Q(1,0) = \frac{3}{4} - \frac{3}{4} + 1 = 1$

$Q(0,1) = 0.02 + 1 = 1.02$

$Q(0.5,0.5) = 0.01 + 1 = 1.01$

$Q(0,1) > Q(0.5,0.5) > Q(1,0)$

$\|\ \|_1$ regularization is Compressive

Regularized objective: $Q(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \|\mathbf{w}\|_1 - \frac{3}{4}$

$Q(1,0) = \frac{3}{4} - \frac{3}{4} + 1 = 1$

$Q(0,1) = 0.02 + 1 = 1.02$

$Q(0.5,0.5) = 0.01 + 1 = 1.01$

$Q(0,1) > Q(0.5,0.5) > Q(1,0)$

$\|\ \|_1$ regularization is Compressive

$Q_2(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \|\mathbf{w}\|^2 - \frac{3}{4}$ $\quad Q_2(1,0) = 1 \quad Q_2(0,1) = 1.02$

Regularized objective: $Q(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \|\mathbf{w}\|_1 - \frac{3}{4}$

$Q(1,0) = \frac{3}{4} - \frac{3}{4} + 1 = 1$

$Q(0,1) = 0.02 + 1 = 1.02$

$Q(0.5,0.5) = 0.01 + 1 = 1.01$

$Q(0,1) > Q(0.5,0.5) > Q(1,0)$

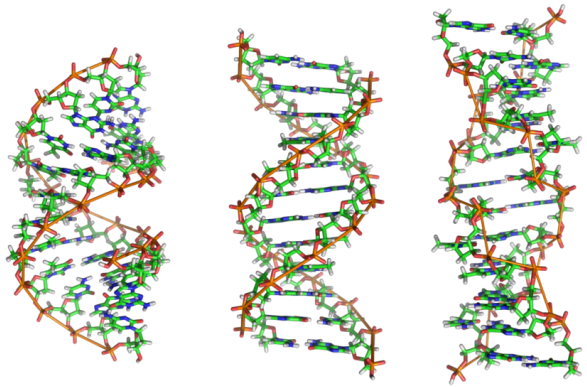
$\|\ \|_1$ regularization is Compressive

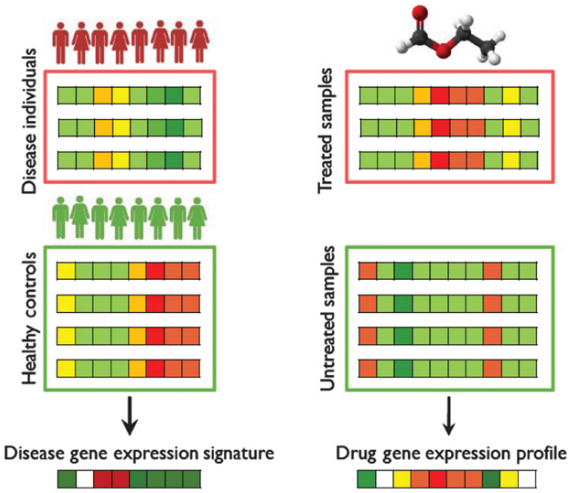
$Q_2(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \|\mathbf{w}\|^2 - \frac{3}{4}$ $\quad Q_2(1,0) = 1 \quad Q_2(0,1) = 1.02$

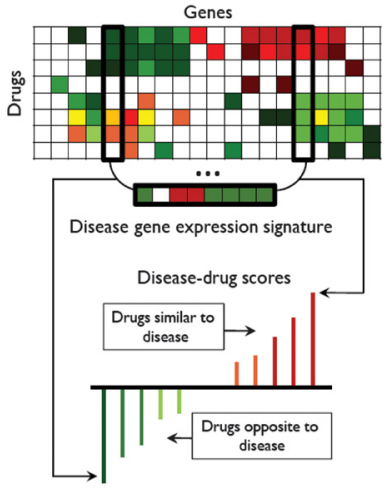
$Q_2(0.5,0.5) = \mathbf{0.51}$

	$\|\mathbf{w}\|_2^2$	$\|\mathbf{w}\|_1$
Computation	Scaling	Shrinkage
Promotes	Diversity	Sparsity
n >> d	✓✓✓	✗
n ~ d	✓✓	✓✓
d >> n	✓	✓✓✓

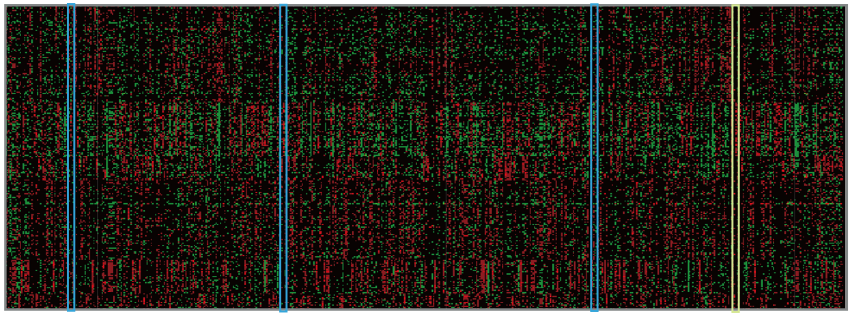
Gene Expression





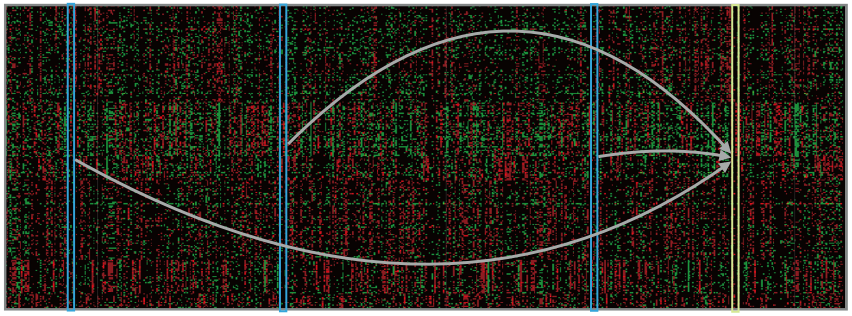


Experimental Expression Matrix



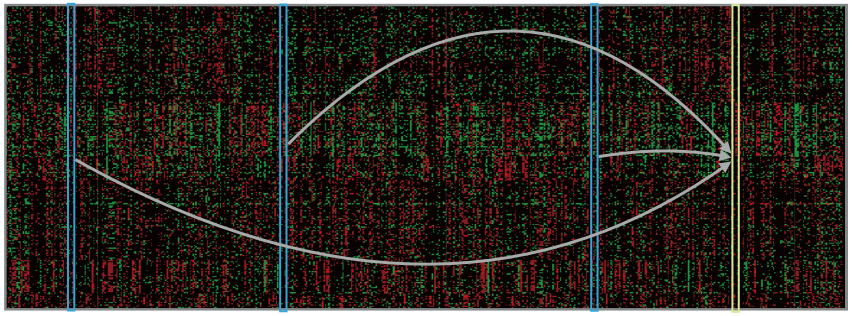
Find few (#genes ≪ #drugs)
highly predictive genes

Experimental Expression Matrix



Find few (#genes ≪ #drugs)
highly predictive genes

Experimental Expression Matrix



Find few (#genes ≪ #drugs)
highly predictive genes

$$\mathcal{Q}(\mathbf{a}) = \mathscr{L}(\mathbf{a}) + \lambda\|\mathbf{a}\|_1 = \left\| \sum_j a_j \mathbf{c}_j - \mathbf{c}_r \right\|^2 + \lambda\|\mathbf{a}\|_1$$


```python
# GD with L1 & L2 regularization
def proxgd(X, eta, l1, l2):
  w = np.zeros((X.shape[1], 1))
  ls = []
  for e in range(len(eta)):
    et = eta[e]
    z = X @ w
    ls.append(loss(z, w, l1, l2))
    q = 1 / (1 + np.exp(-z))
    w = w - et * np.mean(X * q, axis=0, keepdims=True).T
    if l1 > 0:
      w = np.sign(w) * np.maximum(abs(w) - et * l1, 0)
    if l2 > 0:
      w = w / (1 + et * l2)
  return w, ls
```