Regret Bounds for Prediction Problems

Geoff Gordon

ggordon@cs.cmu.edu

September 27, 1999

A simple learning problem

Sequence of numbers $z_1 \ldots z_T$ revealed one at a time

Predict each one—w_t is prediction on trial t

First pick w_1, then see z_1, then pick w_2, then see z_2, ...

Total loss is $\frac{1}{2}\sum_t (w_t - z_t)^2$

Another simple learning problem

N experts predict: will it rain tomorrow, next day, next day, ...

Pick one to believe each day

Can randomize: choose $w_t \geq 0$ with $\sum_i w_{t,i} = 1$

Called "learning from expert advice"

Expert advice cont'd

Loss of expert i on trial t is $x_{t,i} \in \{0, 1\}$ $\quad 1 \le i \le N, 1 \le t \le T$

First pick w_1, then see x_1, then pick w_2 . . .

Expected total loss is $\sum_t w_t \cdot x_t$

Yet another

Linear regression

Regression coefficients w_t, input variables x_t, target output y_t

Loss on trial t is $\frac{1}{2}(y_t - w_t \cdot x_t)^2$

First pick w_1, then see x_1 and y_1, then pick w_2 ...

In common

Trials $1 \ldots T$

Choose prediction w_t

Reveal convex loss function l_t

Minimize $\sum_{t=1}^{T} l_t(w_t)$

Comments

l_t encodes error criterion and training example

We know all of l_t, not just its value at w_t

Includes many standard problems

Can use log likelihood as l_t

What to do?

For squared error from $z_1 \dots z_T$: $w_t =$ mean of z_i, $i < t$

For expert advice: WM

For linear regression: least squares

Inference algorithm

Pick prior loss function l_0

Define $L_t = \sum_{i=0}^{t-1} l_i$

Predict $w_t \in \arg\min_w L_t(w)$

Called MAP

The Weighted Majority algorithm

Start all experts with equal weight

If expert i makes a mistake, multiply weight by $\beta < 1$

Always keep weights normalized so sum is 1

WM is a MAP

Set $l_0(w) = \frac{1}{-\ln\beta}\sum_i (w_i \ln w_i - w_i) + \begin{cases} 0 & w \geq 0, \sum_i w_i = 1 \\ \infty & \text{otherwise} \end{cases}$

MAP says find $\arg\min_w \left[l_0(w) + \sum_t x_t \cdot w\right]$

Second term of l_0 ensures w is a probability distribution

First term ensures

$$w_i \propto \beta^{\sum_t x_{t,i}}$$

since

$$\frac{d}{dw_i}(\text{first term}) \propto \ln w_i$$

Goal

Same problem, same algorithm—why not analyze together?

Will find worst-case regret bound

Use theory of convex duality

Assumptions

Don't want standard statistical assumptions

e.g., independence, identical distributions, known family

but it's impossible to do well in worst case

Regret bounds

Fix $\mathcal{U} \subseteq \mathcal{W}$ ("comparison class")

Compare to best constant prediction $u \in \mathcal{U}$

Hope to do well if $\exists$ a good predictor $u \in \mathcal{U}$

Regret bounds cont'd

Will prove bounds like

$$\sum_{t=1}^{T} l_t(w_t) \leq k_1 \sum_{t=1}^{T} l_t(u) + k_2 \mathsf{size}(\mathcal{U})$$

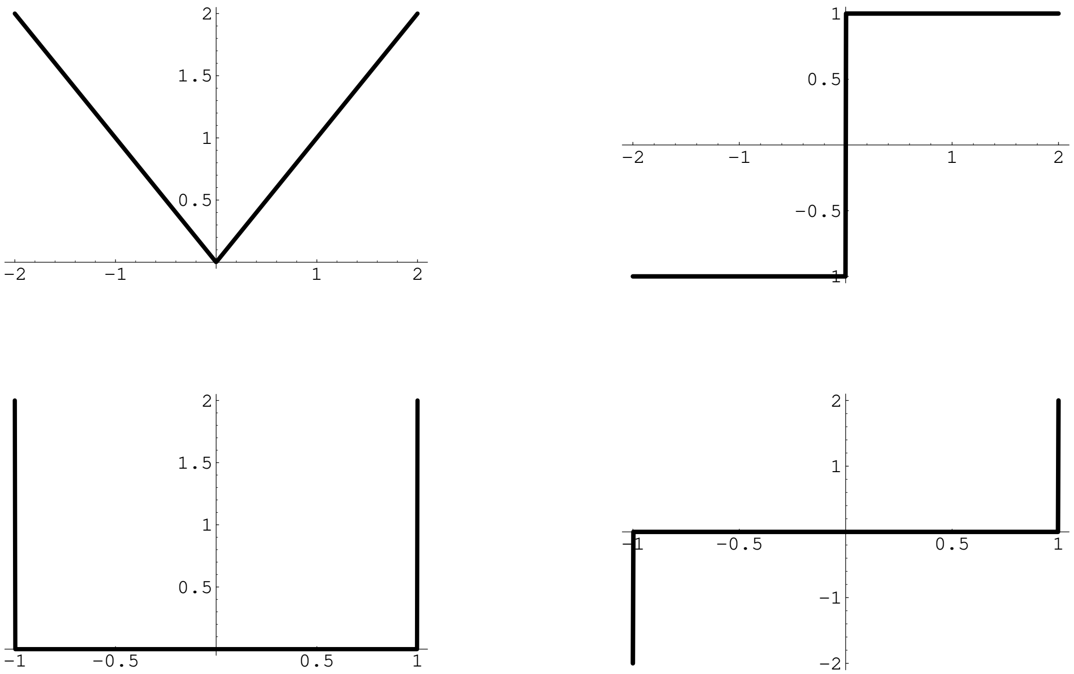
for all $u \in \mathcal{U}$

Regret bounds cont'd

- Example: min squared error from $z_1 \ldots z_T$
- Comparison class might be $\mathcal{U} = \{u \mid ||u||_2 \leq 5\}$
- Best u is mean of $0, z_1, \ldots, z_T$ if in $\mathcal{U}$
- Performance of u is proportional to variance of $0, z_1, \ldots, z_T$
- So, do OK if mean and variance of $z_1, \ldots, z_T$ is small
- Much more general than z_t being $N(\mu, \sigma^2)$ for small μ, σ

Subgradients

- $\partial f(x)$ is set of slopes of tangent planes to f at x
- Write f' for any function with $f'(x) \in \partial f(x)$



Convex duality

To every closed convex f corresponds a unique f^*

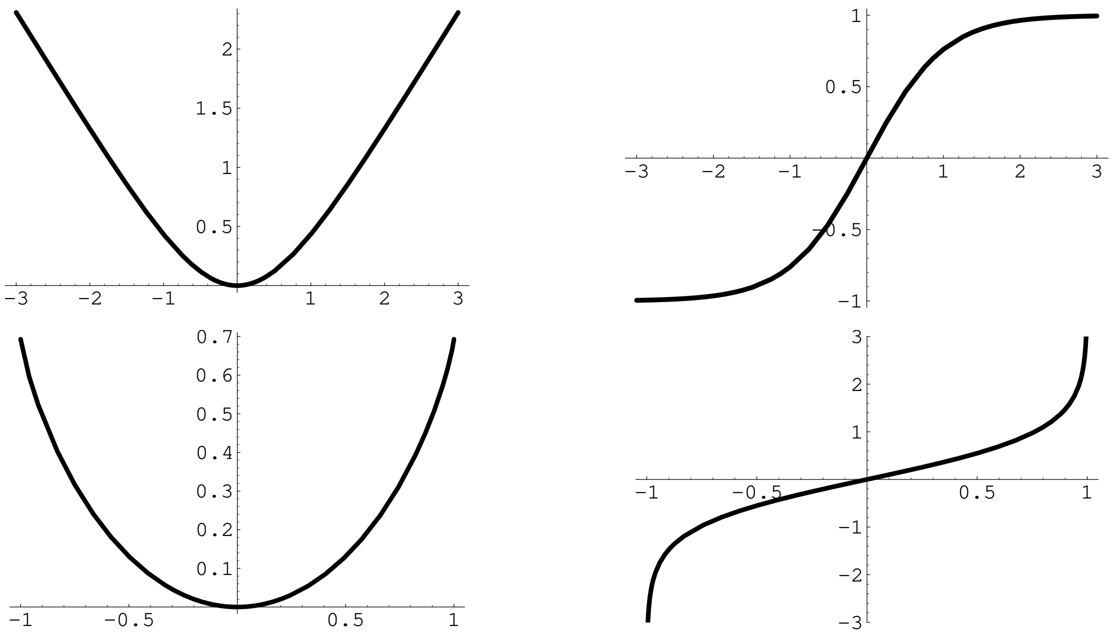
Defined by

$$f^*(y) = \sup_x (y \cdot x - f(x))$$

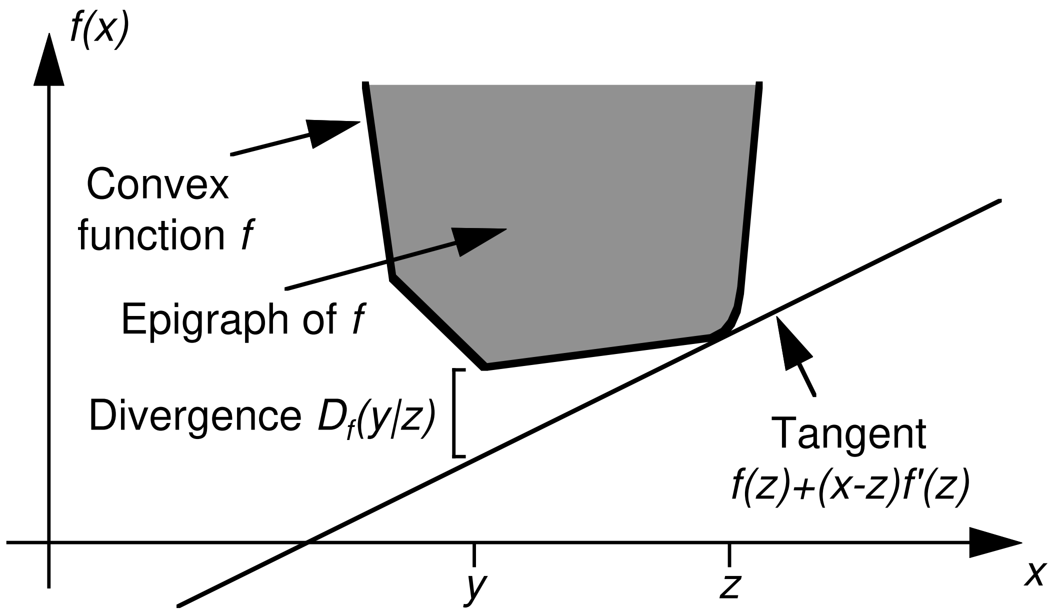
Subgradients are inverses: $a \in \partial f(x) \Leftrightarrow x \in \partial f^*(a)$

related to LP duality, geometric duality, duality of norms, Lagrange multipliers, information geometry

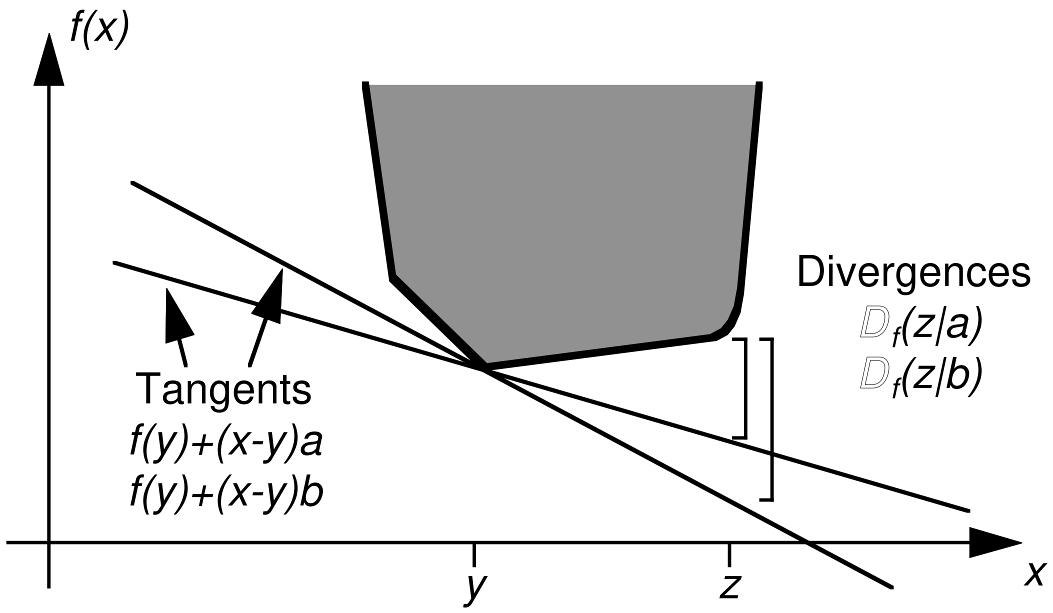
Dual of entropy



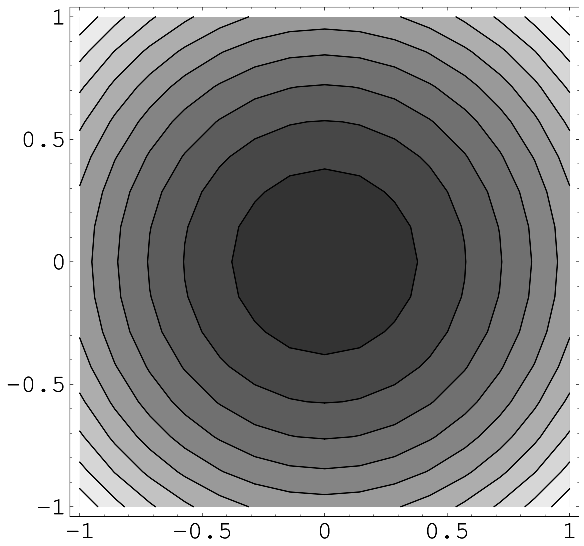
Bregman divergence



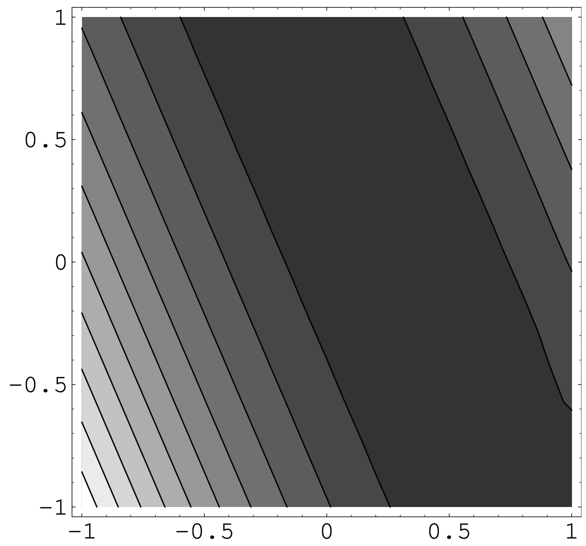
Generalized Bregman divergence



Loss functions are divergences



$$l_0 = w^2$$



$$l_t = (y_t - x_t \cdot w)^2$$

Simple implementation

Suppose l_0 differentiable, l_0' invertible

Suppose l_t linear for $t \geq 1$ (say $l_t(w) = x_t \cdot w$)

Minimum on trial t is at

$$\begin{aligned} 0 &= \frac{d}{dw}\left(l_0(w) + \sum_{i=1}^{t-1} x_i \cdot w\right) \\ &= l_0'(w) + \sum_{i=1}^{t-1} x_i \\ w &= (l_0')^{-1}(X_t) \end{aligned}$$

Implementation cont'd

To find prediction of MAP, keep sum of $-x_t$ and apply $(l'_0)^{-1}$

Same is true for any convex l_0

- derivatives $\rightarrow$ subgradients
- $(l'_0)^{-1} \rightarrow \partial l^*_0$

∂l^*_0 is called a link function (*cf* GLIM)

Proof sketch

Amortized analysis

Potential is a Bregman divergence related to loss functions

Required assumption: l_t not too convex compared to $\sum_{i=0}^{t-1} l_i$

The bounds

$$\sum_{t=1}^{T} l_t(w_t) \leq \sum_{t=1}^{T} \frac{1}{1-\eta_t} l_t(u) + \frac{1}{1-\eta_1} U$$

$\frac{1}{\eta_t}$ measures how much information is available on trial t

U measures size of comparison class $\mathcal{U}$

Summary

Provided a unified framework for reasoning about worst-case regret bounds

Placed several known algorithms within this framework

Read the paper

More detail in my COLT99 paper

Even more (and more recent) detail in my thesis

Both available from

www.cs.cmu.edu/~ggordon

Topics not covered

Gradient descent: use linear lower bound on true loss function

Inference in exponential families:

- MAP algorithm is to keep average of sufficient statistics
- connection between links/divergences, information geometry

Topics not covered, cont'd

Link function:

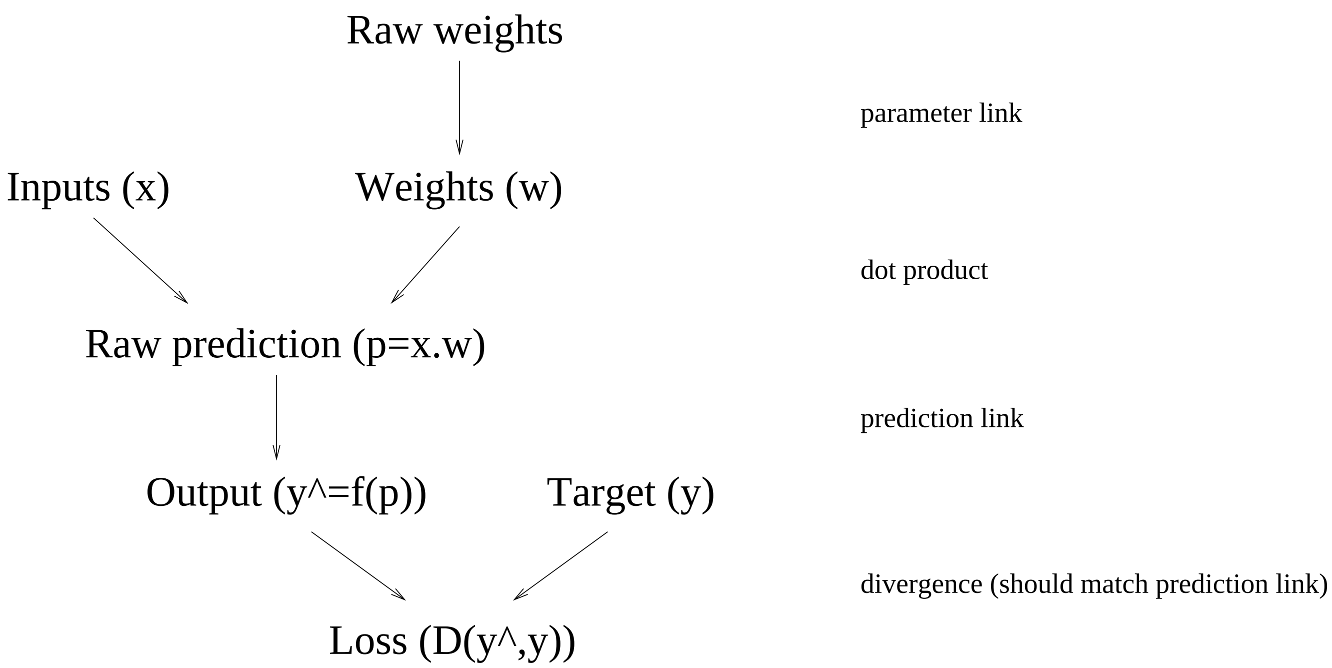
- subgradient of loss
- connects two spaces (primal and dual, point and tangent, expectation parameters and natural parameters)
- may need several links connecting several spaces

Regression problems:

- two separate link funs
- matching losses for nondifferentiable links

Slides after this point were not presented during the main part of the talk—I just brought them in case someone asked me the relevant question

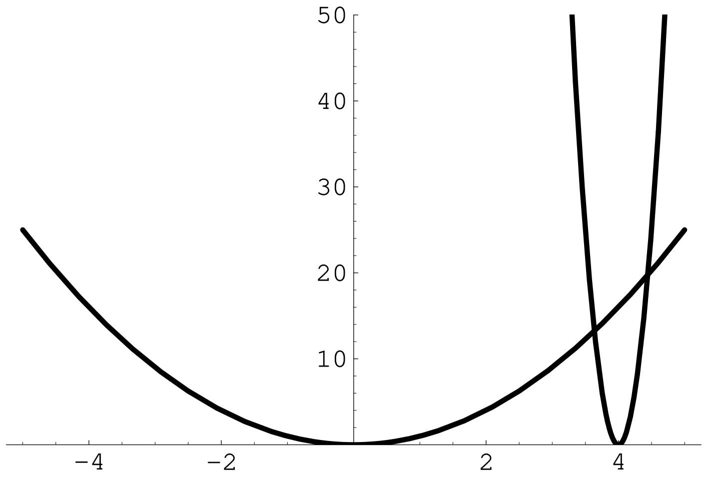
Spaces involved in regression



An example

Trial t	0	1	2	3	4
Prediction w_t	—	0	2	3	3
Training example z_t	—	4	5	3	8
Error type	—	Squared	Squared	Squared	Squared
Loss function $l_t(w)$	w^2	$(w-4)^2$	$(w-5)^2$	$(w-3)^2$	$(w-8)^2$
Loss of w_t	—	16	9	0	25
Ttl loss of $w_1 \dots w_t$	0	16	25	25	50
Best constant u	4				
Loss of u	16	0	1	1	16
Ttl loss of u	16	16	17	18	34
Ttl regret	-16	0	8	7	16

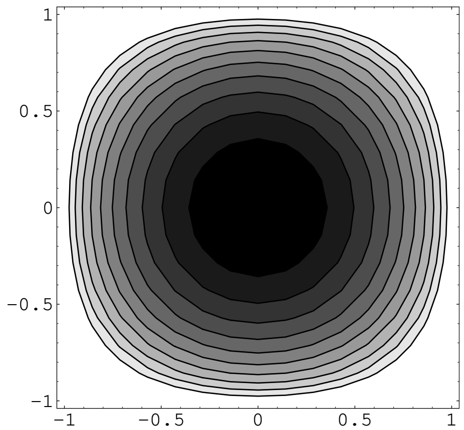
Information



- How well can we do if $\exists$ a good u?
- Depends on amount of information on each trial
- Loosely, too much info if l_t more convex than $\sum_{i=0}^{t-1} l_i$
- think Fisher

Definition of U

- U is bound on $\mathbb{D}_{l_0}(u|0)$
- $\mathcal{U} = \{u|\mathbb{D}_{l_0}(u|0) \leq U\}$



Definition of η_t

$$\eta_t \mathbb{D}_{L_t}(v|a) \geq \frac{1}{2}(g(v-w))^2$$

for all v and w and $a \in \partial L_t(w)$

$$\mathbb{D}_{l_t^*}(0|w) \geq \frac{1}{2}(g^\circ(a))^2$$

for all w and $a \in \partial l_t(w)$

Definition of η_t