Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta

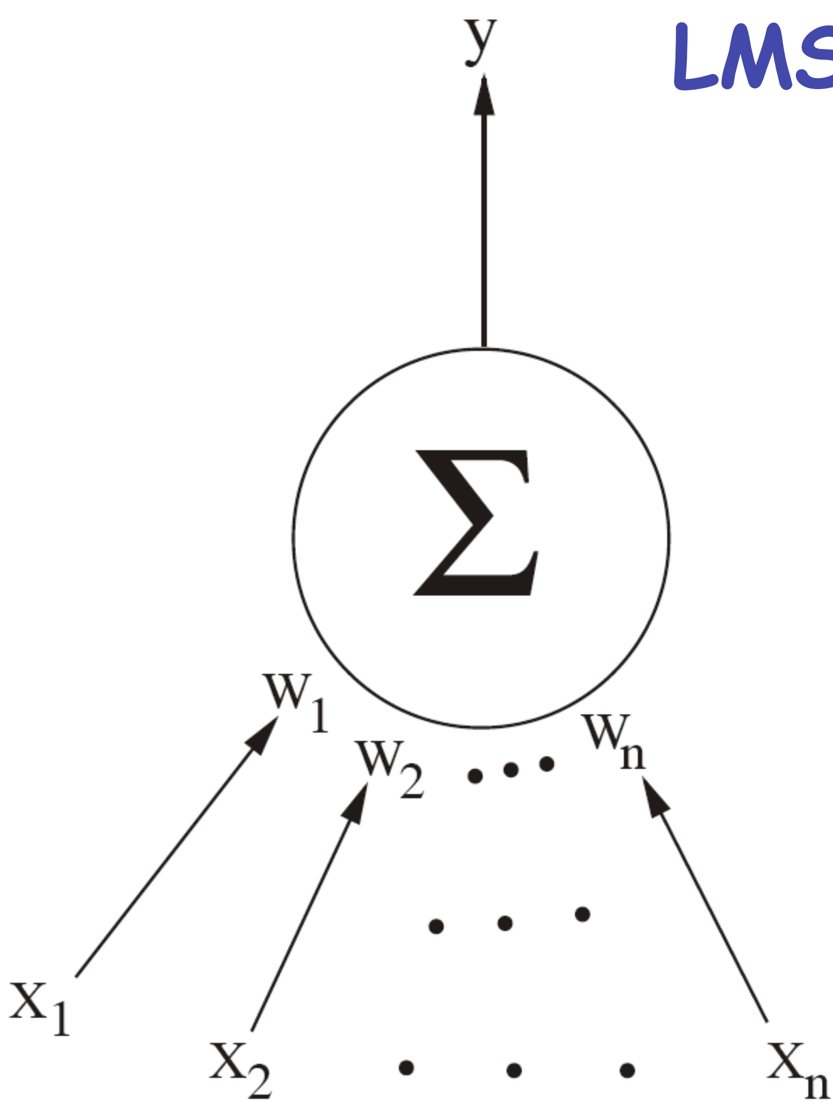
Rich Sutton

in AAAI-92

Outline

- Remarks on learning and bias
- Least Mean Square (LMS) learning
- IDBD (new algorithm)
- Exp 1: Is IDBD better?
- Exp 2: Does IDBD find an optimal bias?
- Derivation of LMS as gradient descent
- Derivation of IDBD as “meta” grad desc

LMS (least mean square) Learning



$$y(t) = \sum_{i=1}^{n} w_i(t) x_i(t)$$

error $\delta(t) = y^*(t) - y(t)$

Minimize expected $\delta^2(t)$

$$w_i(t+1) = w_i(t) + \alpha \delta(t) x_i(t)$$

Incremental Delta-Bar-Delta (IDBD)

$$w_i(t+1) = w_i(t) + \alpha_i(t+1)\delta(t)x_i(t)$$

$$\alpha_i(t) = e^{\beta_i(t)}$$

$$\beta_i(t+1) = \beta_i(t) + \theta\delta(t)x_i(t)h_i(t)$$

$$h_i(t+1) = h_i(t)\Big[1 - \alpha_i(t+1)x_i^2(t)\Big]^+ + \alpha_i(t+1)\delta(t)x_i(t)$$

where $[x]^+$ is x, if $x > 0$, else 0

Experiment 1: Does IDBD Help?

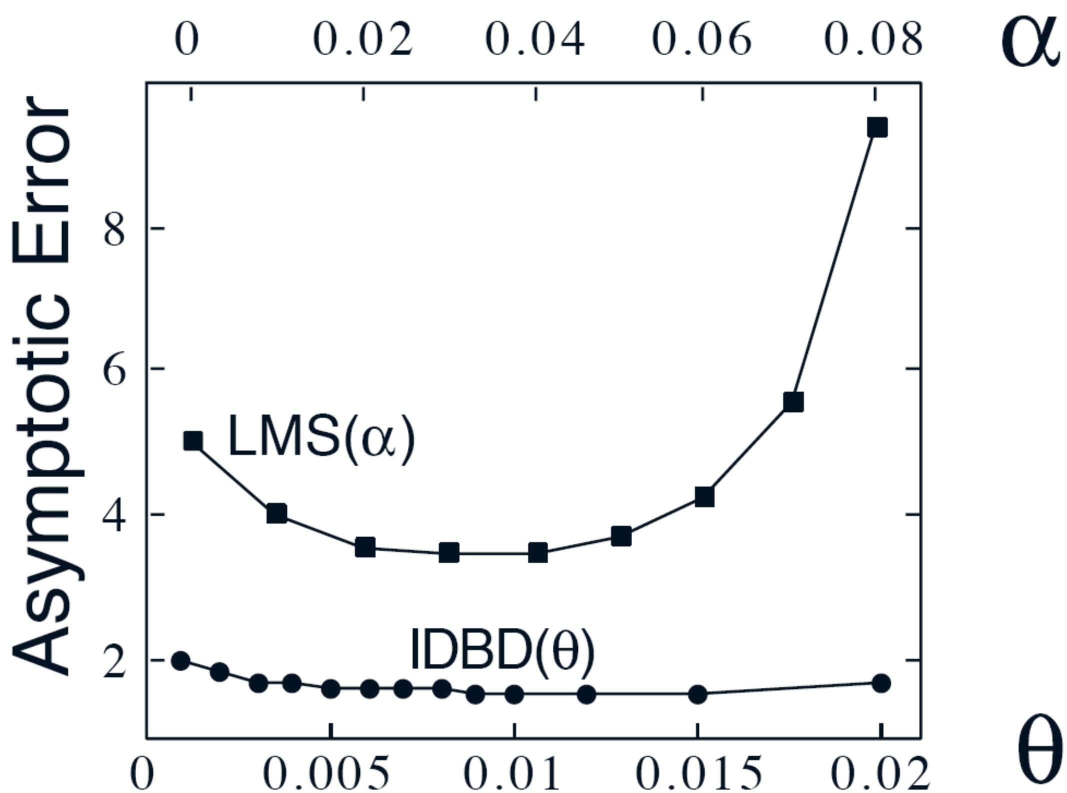
20 inputs,
5 relevant
$s_i \in \{-1, +1\}$

$$y^* = s_1x_1 + s_2x_2 + s_3x_3 + s_4x_4 + s_5x_5 + 0x_6 + 0x_7 + \cdots + 0x_{20},$$

one s_i switched
every 20 examples

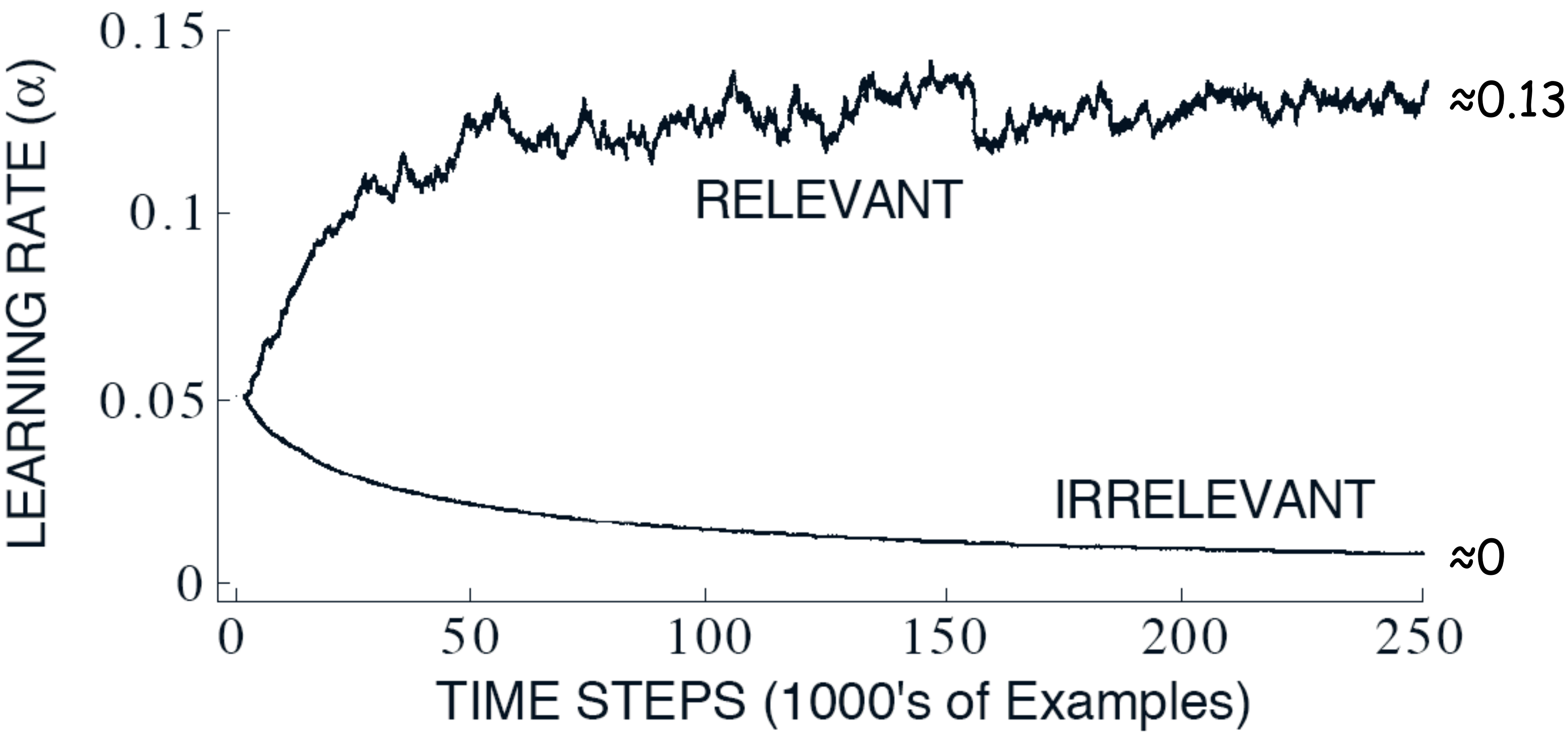
20,000 examples
to wash transients,
then 10,000 examples
for real

Average MSE
measured over the
10,000 examples



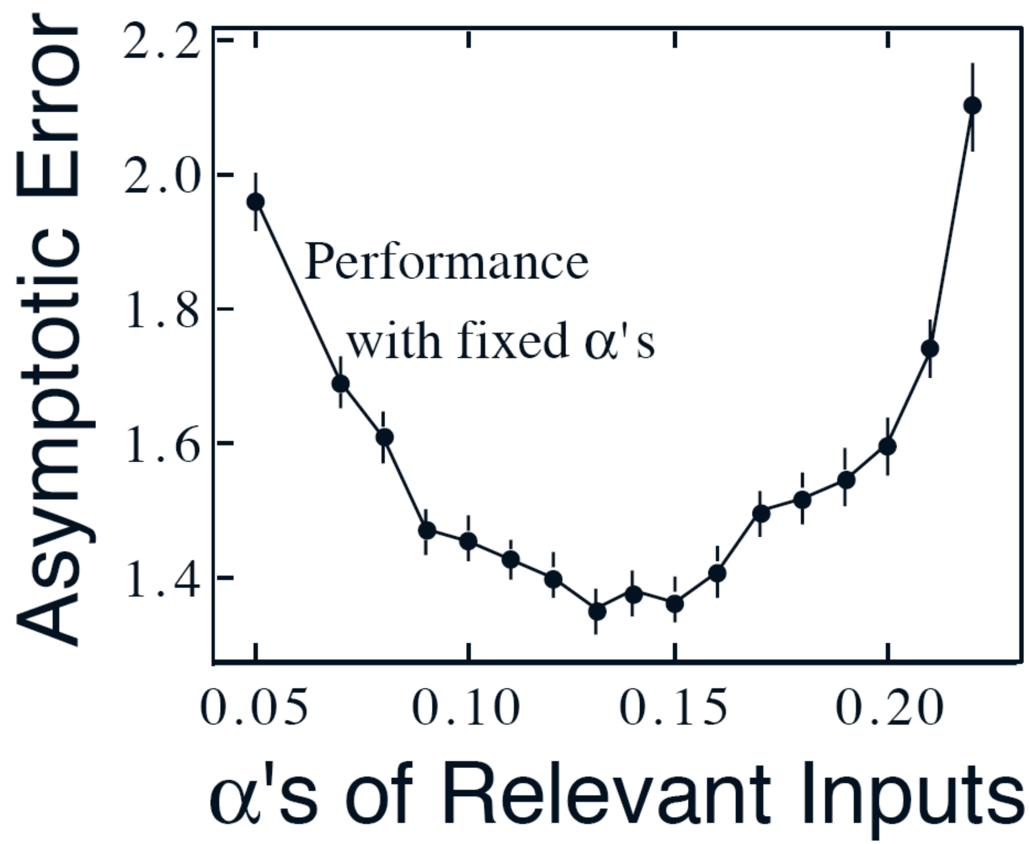
Exp 2a: What αs are found by IDBD?

Small θ (=0.001), long run



Exp 2b: What are the optimal αs?

Repeat of first experiment with fixed αs: α=0 for irrelevant inputs, vary α for relevant inputs



Optimal $\alpha \approx 0.13$

Derivation of LMS as grad descent (1)

Error is $\delta(t) = y^*(t) - y(t)$

Seek to minimize expected squared error

using the sample squared error $\delta^2(t)$

Ergo, the gradient descent idea:

$$w_i(t+1) = w_i(t) - \frac{1}{2}\alpha \frac{\partial \delta^2(t)}{\partial w_i(t)}$$

Derivation of LMS as grad descent (2)

$$
\begin{aligned}
w_i(t+1) &= w_i(t) - \frac{1}{2}\alpha \frac{\partial \delta^2(t)}{\partial w_i(t)} \\
&= w_i(t) - \alpha\delta(t)\frac{\partial \delta(t)}{\partial w_i(t)} \\
&= w_i(t) - \alpha\delta(t)\frac{\partial [y^*(t) - y(t)]}{\partial w_i(t)} \\
&= w_i(t) + \alpha\delta(t)\frac{\partial y(t)}{\partial w_i(t)} \\
&= w_i(t) + \alpha\delta(t)\frac{\partial}{\partial w_i(t)}\Big[\sum_{j=1}^{n} w_j(t)x_j(t)\Big] \\
&= w_i(t) + \alpha\delta(t)x_i(t) \quad ✔
\end{aligned}
$$

Derivation of IDBD as GD (1)

$$\beta_i(t+1) = \beta_i(t) - \frac{1}{2}\theta \frac{\partial \delta^2(t)}{\partial \beta_i}$$

$$= \quad \beta_i(t) - \frac{1}{2}\theta \sum_j \frac{\partial \delta^2(t)}{\partial w_j(t)} \frac{\partial w_j(t)}{\partial \beta_i}$$

assuming

$$\frac{\partial w_j(t)}{\partial \beta_i} \approx 0$$

for $i \neq j$

$$\approx \quad \beta_i(t) - \frac{1}{2}\theta \frac{\partial \delta^2(t)}{\partial w_i(t)} \frac{\partial w_i(t)}{\partial \beta_i}$$

$$= \quad \beta_i(t) + \theta \delta(t) x_i(t) h_i(t)$$ ✔

where $h_i(t) = \dfrac{\partial w_i(t)}{\partial \beta_i}$

Derivation of IDBD as GD (2)

$$
\begin{aligned}
\frac{\partial \delta(t)}{\partial \beta_i} &= -\frac{\partial y(t)}{\partial \beta_i} = -\frac{\partial}{\partial \beta_i} \sum_j w_j(t) x_j(t) \\
&\approx -\frac{\partial}{\partial \beta_i} \Big[w_i(t) x_i(t) \Big] = -h_i(t) x_i(t)
\end{aligned}
$$

$$
\begin{aligned}
h_i(t+1) &= \frac{\partial w_i(t+1)}{\partial \beta_i} \\
&= \frac{\partial}{\partial \beta_i} \Big[w_i(t) + e^{\beta_i(t+1)} \delta(t) x_i(t) \Big] \\
&= h_i(t) + e^{\beta_i(t+1)} \delta(t) x_i(t) + e^{\beta_i(t+1)} \frac{\partial \delta(t)}{\partial \beta_i} x_i(t) \\
&\approx h_i(t) + e^{\beta_i(t+1)} \delta(t) x_i(t) - e^{\beta_i(t+1)} x_i^2(t) h_i(t) \\
&= h_i(t) \Big[1 - \alpha_i(t+1) x_i^2(t) \Big] + \alpha_i(t+1) \delta(t) x_i(t) \ ✔
\end{aligned}
$$

Closing

- IDBD improves over earlier DBD methods
  - Incremental, example-by-example updating
  - Only one free parameter
  - Can be derived as "meta" gradient descent
- IDBD is a principled way of learning bias (for linear LMS methods)
  - As such can dramatically speed learning
  - With only a linear (≈x3) increase in memory & comp.
- "Meta" gradient descent is a general idea
  - An incremental form of hold-one-out cross validation
  - Fundamentally different from other learning methods
  - Could be used to learn other kinds of biases