Reparameterizing Mirror Descent as Gradient Descent

Ehsan Amid and Manfred K. Warmuth

Google Research, Brain Team

{eamid, manfred}@google.com

Mirror descent

$$\boldsymbol{w}_{s+1} = f^{-1}(f(\boldsymbol{w}_s) - h\,\nabla L(\boldsymbol{w}_s))$$

where f is (coordinate-wise) strictly monotonic link function

Gradient Descent (GD):

$$\boldsymbol{w}_{s+1} = \boldsymbol{w}_s - h\,\nabla L(\boldsymbol{w}_s) \qquad (f(\boldsymbol{w}) = \boldsymbol{w})$$

Unnormalized Exponentiated Gradient Descent (EGU): [KW97]

$$\boldsymbol{w}_{s+1} = \boldsymbol{w}_s \odot \exp(-h\,(\nabla L(\boldsymbol{w}_s)) \qquad (f(\boldsymbol{w}) = \log \boldsymbol{w})$$

(with $w_i \geq 0$)

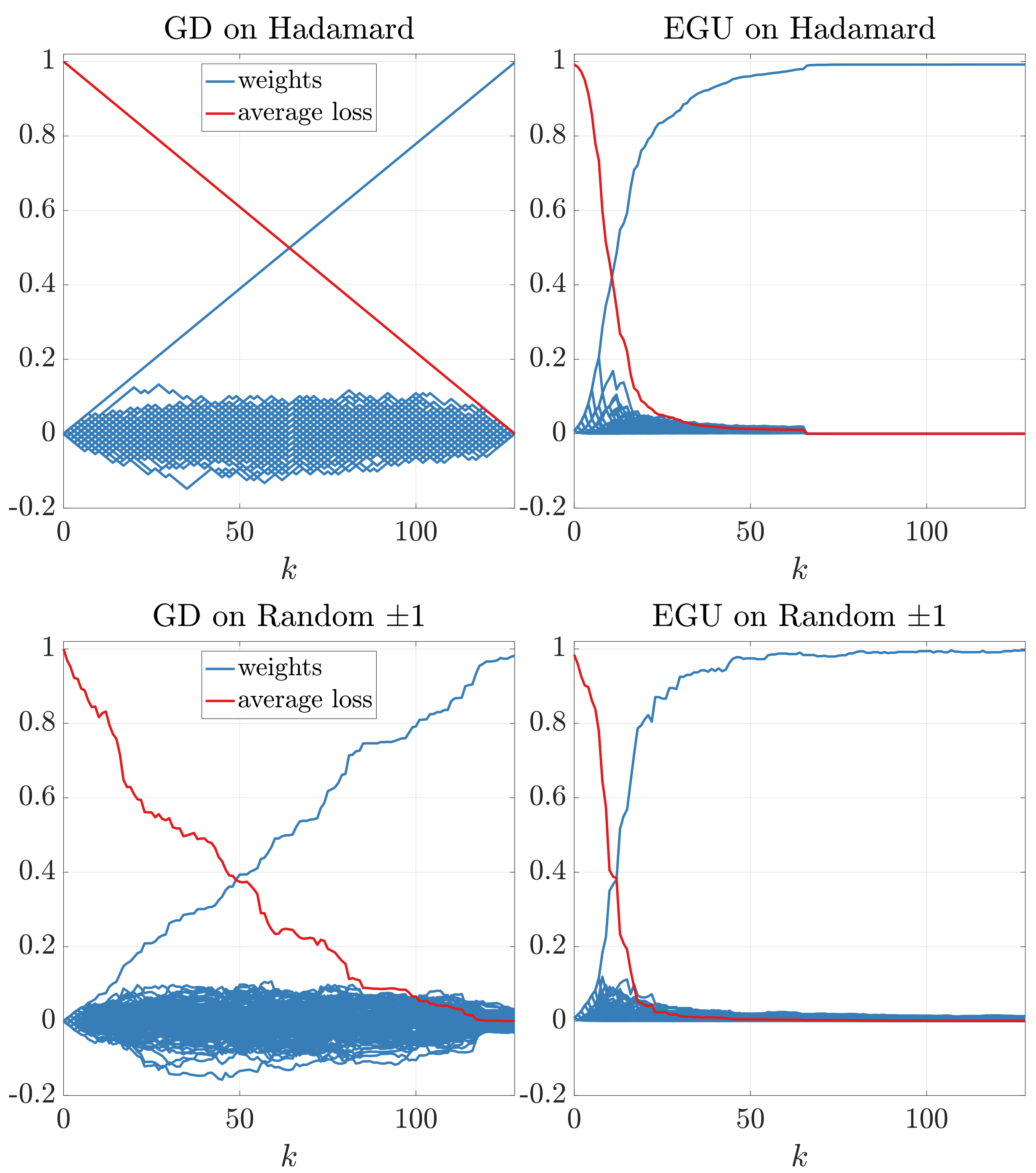
Major differences

GD: backprop, kernel methods

EG: Winnow, expert algorithms, Boosting, Bayes

Setup: **128x128 Hadamard (top) and random (bottom) ± 1 matrix**

Rows are instances, labels are the first column, square loss



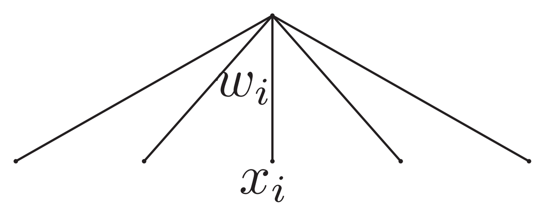
x-axis: $k = 1..128$

y-axis: all 128 weights Loss when trained on examples $1..k$

Upshot: After half examples, GD has average loss $1/2$

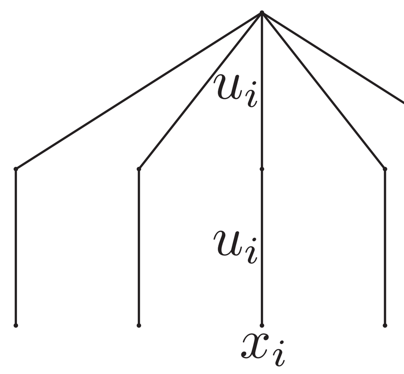
EG family converges in $\mathcal{O}(\log(n))$ many examples

Surprises



When linear neuron is trained with GD: lower bound for linear decrease of av. loss [WV05]

Reparameterize weights w_i by u_i^2 [Akin79,GWBNS17]



Continuous GD on u_i simulates continuous-time EGU on w_i

Discretizations learn Hadamard with Backprop with essentially $\mathcal{O}(\log n)$ examples

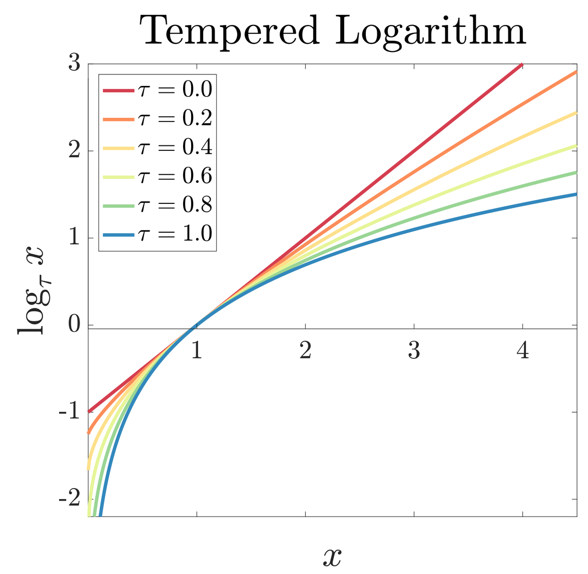
Experimentally indistinguishable from discretized EGU

Focus on continuous MD

$$\dot{f}(\boldsymbol{w}) = -\nabla L(\boldsymbol{w})$$

Main examples:
GD ($f(\boldsymbol{w}) = \boldsymbol{w}$) and EGU ($f(\boldsymbol{w}) = \log(\boldsymbol{w})$)

Between $f(\boldsymbol{w}) = \boldsymbol{w}$ & $f(\boldsymbol{w}) = \log \boldsymbol{w}$:
$\log_\tau(\boldsymbol{w}) := \frac{1}{1-\tau}(\boldsymbol{w}^{1-\tau} - 1)$
(for $\tau \in [0, 1]$)



Reparameterization

Main theorem: For the reparameterization function $\boldsymbol{w} = q(\boldsymbol{u})$ with the property that range(q) = dom(f), the two updates

$$\dot{f}(\boldsymbol{w}) = -\nabla L(\boldsymbol{w}) \quad \text{and} \quad \dot{\boldsymbol{u}} = -\nabla L \circ q\,(\boldsymbol{u}),$$

coincide if that $\boldsymbol{w}(0) = q(\boldsymbol{u}(0))$, range$(q) \subseteq$ dom(F), and we have

$$(\boldsymbol{J}_f(\boldsymbol{w}))^{-1} = \boldsymbol{J}_q(\boldsymbol{u})\;(\boldsymbol{J}_q(\boldsymbol{u}))^\top$$

Main case: EGU as GD

Link

$$f(\boldsymbol{w}) = \log(\boldsymbol{w})$$

Reparameterization

$$\boldsymbol{w} = q(\boldsymbol{u}) := {}^1\!/\!_4\,\boldsymbol{u} \odot \boldsymbol{u}$$

$$(\boldsymbol{J}_f(\boldsymbol{w}))^{-1} = (\mathrm{diag}(\boldsymbol{w})^{-1})^{-1} = \mathrm{diag}(\boldsymbol{w})$$

$$\boldsymbol{J}_q(\boldsymbol{u})(\boldsymbol{J}_q(\boldsymbol{u}))^\top = {}^1\!/\!_2\,\mathrm{diag}(\boldsymbol{u})\;({}^1\!/\!_2\,\mathrm{diag}(\boldsymbol{u}))^\top = \mathrm{diag}(\boldsymbol{w})$$

Therefore

$$\dot{\log}(\boldsymbol{w}) = -\nabla L(\boldsymbol{w}) \quad \text{equals} \quad \dot{\boldsymbol{u}} = -\underbrace{\nabla L \circ q\,(\boldsymbol{u})}_{\nabla_{\boldsymbol{u}} L\,({}^1\!/\!_4\,\boldsymbol{u} \odot \boldsymbol{u})}$$

Burg as GD

$$(\dot{-\boldsymbol{1} \oslash \boldsymbol{w}}) = -\nabla L(\boldsymbol{w}) \quad \text{equals} \quad \dot{\boldsymbol{u}} = -\underbrace{\nabla L \circ q\,(\boldsymbol{u})}_{\nabla_{\boldsymbol{u}} L\,(\exp(\boldsymbol{u}))}$$

Tempered $\log_\tau$ as GD

$$\dot{\log_\tau}(\boldsymbol{w}) = -\nabla L(\boldsymbol{w}) \quad \text{equals} \quad \dot{\boldsymbol{u}} = -\underbrace{\nabla L \circ q\,(\boldsymbol{u})}_{\nabla_{\boldsymbol{u}} L\,\left(\left(\frac{2-\tau}{2}\right)^{\frac{2}{2-\tau}} \boldsymbol{u}^{\frac{2}{2-\tau}}\right)}$$

Conclusion

- World of continuous updates more succinct

  Euler discr.: $$\frac{f(\boldsymbol{w}(t+h)) - f(\boldsymbol{w}(t))}{h} = -\nabla L(\boldsymbol{w}(t))$$
  $$\iff \boldsymbol{w}(t+h) = f^{-1}(f(\boldsymbol{w}(t)) - h\nabla L(\boldsymbol{w}(t)))$$

- Under what conditions does the discrete MD track continuous MD
- Discretization of reparameterized EGU as GD tracks discrete EGU well enough so that same regret bounds hold [AW20]
- Discretization of reparameterized EGU as GD sample efficiently learns Hadamard problem