Reparameterizing Mirror Descent as Gradient Descent

Ehsan Amid & Manfred K. Warmuth
Google Research, Brain Team
{eamid, manfred}@google.com
Mirror descent

$$w_{s+1} = f^{-1}(f(w_s) - h \nabla L(w_s))$$

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

Gradient Descent (GD):

$$w_{s+1} = w_s - h \nabla L(w_s)$$

($f(w) = w$)

Unnormalized Exponentiated Gradient Descent (EGU): [KW97]

$$w_{s+1} = w_s \odot \exp(-h (\nabla L(w_s)))$$

($f(w) = \log w$)

(with $w_i \geq 0$)
Major differences between the two families

GD: backprop, kernel methods
EGU: Winnow, expert algorithms, Boosting, Bayes

Setup: **128x128 Hadamard matrix**
**Permuted** rows are instances, labels are any fixed column

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 $= \frac{1}{2}$
EG family converges in log($n$) many examples
Major differences between the two families

GD: backprop, kernel methods
EG: Winnow, expert algorithms, Boosting, Bayes

Setup: \textbf{128x128 random \pm 1 matrix}
Rows are instances, labels are the first column, square loss

\begin{figure}
\centering
\includegraphics[width=\textwidth]{gd_eg_plot.png}
\caption{GD and EG on Random \pm 1}
\end{figure}

\begin{itemize}
  \item \textbf{x-axis:} \( k = 1..128 \)
  \item \textbf{y-axis:} all 128 weights \textbf{Loss when trained on examples 1..k}
\end{itemize}

Upshot: After half examples, GD has average loss \( \approx \frac{1}{2} \)
EG family converges in log\((n)\) many examples
When linear neuron is trained with GD, then lower bound for linear decrease of avg. loss [WV05]

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

Continuous GD on $u_i$ simulates continuous EGU on $w_i$

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

Experimentally indistinguishable from discrete EGU
Main focus here: continuous MD

\[ \dot{f}(w) = -\nabla L(w) \]

Main examples:
GD \((f(w) = w)\) and EGU \((f(w) = \log(w))\)

Between \(f(w) = \log w\) and \(f(w) = w\):
\[
\log_\tau(w) := \frac{1}{1-\tau}(w^{1-\tau} - 1)
\]
(for \(\tau \in [0, 1]\))
Main Theorem: For the reparameterization function $w = q(u)$ with the property that $\text{range}(q) = \text{dom}(f)$, the two updates

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

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

$$(\mathbf{J}_f(w))^{-1} = \mathbf{J}_q(u) \mathbf{J}_q(u)^\mathsf{T}$$
EGU as GD: The squaring trick

Link

\[ f(\mathbf{w}) = \log(\mathbf{w}) \]

Reparameterization

\[ \mathbf{w} = q(\mathbf{u}) := \frac{1}{4} \mathbf{u} \odot \mathbf{u} \]
\[ \mathbf{u} = 2\sqrt{\mathbf{w}} \]

\[ (J_f(\mathbf{w}))^{-1} = (\text{diag}(\mathbf{w})^{-1})^{-1} = \text{diag}(\mathbf{w}) \]
\[ J_q(\mathbf{u})(J_q(\mathbf{u}))^\top = \frac{1}{2} \text{diag}(\mathbf{u}) (\frac{1}{2} \text{diag}(\mathbf{u}))^\top = \text{diag}(\mathbf{w}) \]

Conclusion

\[ \dot{\log}(\mathbf{w}) = -\nabla L(\mathbf{w}) \quad \text{equals} \quad \dot{\mathbf{u}} = -\nabla \circ q(\mathbf{u}) \]
\[ \nabla u L \left( \frac{1}{4} \mathbf{u} \odot \mathbf{u} \right) \]
Burg as GD

Link

\[ f(w) = -1 \odot w \]

Reparameterization

\[ w = q(u) := \exp(u) \]
\[ u = \log(w) \]

\[
(J_f(w))^{-1} = \text{diag}(1 \odot (w \odot w))^{-1} = \text{diag}(w)^2
\]

\[
J_q(u)(J_q(u))^\top = \text{diag}(\exp(u)) \text{diag}(\exp(u))^\top = \text{diag}(w)^2
\]

Conclusion

\[
(-1 \odot w) = -\nabla L(w) \quad \text{equals} \quad \dot{u} = -\nabla L \circ q(u)
\]

\[
\dot{u} = \nabla_u L(\exp(u))
\]
\[ \log_\tau \mathbf{w} = \frac{1}{1-\tau} (\mathbf{w}^{1-\tau} - 1) \] as GD

Link

\[ f(\mathbf{w}) = \log_\tau \mathbf{w} \]

Reparameterization

\[ \mathbf{w} = q(\mathbf{u}) := \left( \frac{2 - \tau}{2} \right)^{\frac{2}{2-\tau}} \mathbf{u}^{\frac{2}{2-\tau}} \]

\[ \mathbf{u} = \frac{2}{2-\tau} \mathbf{w}^{\frac{2}{2-\tau}} \]

\[ (J_{\log_\tau (\mathbf{w})})^{-1} = (\text{diag}(\mathbf{w})^{-\tau})^{-1} = \text{diag}(\mathbf{w})^\tau \]

\[ J_q(\mathbf{u})(J_q(\mathbf{u}))^\top = \left( \left( \frac{2 - \tau}{2} \right)^{\frac{\tau}{2-\tau}} \text{diag}(\mathbf{u})^{\frac{\tau}{2-\tau}} \right)^2 = \text{diag}(\mathbf{w})^\tau \]

Conclusion

\[ \log_\tau (\mathbf{w}) = -\nabla L(\mathbf{w}) \] equals \[ \dot{\mathbf{u}} = -\nabla L \circ q(\mathbf{u}) \]

\[ \nabla_u L \left( \left( \frac{2 - \tau}{2} \right)^{\frac{2}{2-\tau}} \mathbf{u}^{\frac{2}{2-\tau}} \right) \]
Open problems

- World of continuous updates more succinct

\[
\text{Euler discr.: } \frac{f(w(t + h)) - f(w(t))}{h} = -\nabla L(w(t))
\]

\[\iff \quad w(t + h) = f^{-1}(f(w(t))) - h \nabla L(w(t)))\]

- Under what conditions does the discrete MD track continuous MD

Discretization of reparameterized EGU as GD tracks discrete EGU well enough so that the same regret bounds hold \[\text{[AW20]}\]

Discretization of reparameterized EGU as GD sample efficiently learns Hadamard problem