Competitive Gradient Descent

Florian Schäfer and Anima Anandkumar

Caltech

We gratefully acknowledge support by

Competitive optimization

Two agents choose their decision variables to optimize their own objective. Conflicting objectives depend on both players' actions.

$$\min_x f(x,y), \qquad \min_y g(x,y)$$

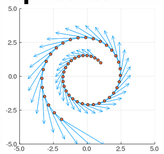
Arises in constrained optimization, robust statistics, and ML (GANs).

Rock! Paper! Scissor! Rock! Paper! …

Naïve: Simultaneous Gradient Descent

$$x_{k+1} = x_k - \eta \nabla_x f(x_k, y_k),$$
$$y_{k+1} = y_k - \eta \nabla_y g(x_k, y_k)$$

Divergent behavior even for simple bilinear game $f(x,y) = -g(x,y) = xy$!



https://commons.wikimedia.org/wiki/File:Rock-paper-scissors.svg

What is gradient descent in two-player games?

Gradient descent minimizes quadratically regularized first order approximation:

$$x_{k+1} = x_k + \operatorname*{argmin}_x f(x_k) + x^T \nabla_x f(x_k) + \frac{x^T x}{2\eta}$$

Thus, the generalization should be obtained from the **Nash equilibrium of a quadratically regularized first order approximation**

Linear or Bilinear approximation?

"Linear for one player ⇒ Linear for two players" loose the interactive aspect:

$$x_{k+1} = x_k + \operatorname*{argmin}_x x^T \nabla_x f + x^T D^2_{xy} f\, y + y^T \nabla_y f + \frac{x^T x}{2\eta}$$

$$y_{k+1} = y_k + \operatorname*{argmin}_y x^T \nabla_x g + x^T D^2_{xy} g\, y + y^T \nabla_y g + \frac{y^T y}{2\eta}$$

All derivatives evaluated in (x_k, y_k)

"Linear for one player ⇒ Bilinear for two players" leads to interactive local game!

The local game has a unique Nash equilibrium

Theorem: The local game has a unique Nash equilibrium given by

$$x = -\eta\left(\mathrm{Id} - \eta^2 D^2_{xy} f\, D^2_{yx} g\right)^{-1}\left(\nabla_x f - \eta D^2_{xy} f\, \nabla_y g\right)$$
$$y = -\eta\left(\mathrm{Id} - \eta^2 D^2_{yx} g\, D_{xy} f\right)^{-1}\left(\nabla_y g - \eta D^2_{yx} g\, \nabla_x f\right)$$

Novel algorithm uses local Nash as update rule

Algorithm: [Competitive Gradient Descent (CGD)]:

At each step, compute (x_{k+1}, y_{k+1}) from (x_k, y_k) as

$$x_{k+1} = x_k - \eta\left(\mathrm{Id} - \eta^2 D^2_{xy} f\, D^2_{yx} g\right)^{-1}\left(\nabla_x f - \eta D^2_{xy} f\, \nabla_y g\right)$$
$$y_{k+1} = y_k - \eta\left(\mathrm{Id} - \eta^2 D^2_{yx} g\, D_{xy} f\right)^{-1}\left(\nabla_y g - \eta D^2_{yx} g\, \nabla_x f\right)$$

What I think that they think that I think …

Can write CGD update as

$$\begin{pmatrix} x_{k+1} - x_k \\ y_{k+1} - y_k \end{pmatrix} = -\eta \begin{pmatrix} \mathrm{Id} & \eta D^2_{xy} f \\ \eta D^2_{yx} g & \mathrm{Id} \end{pmatrix}^{-1} \begin{pmatrix} \nabla_x f \\ \nabla_y g \end{pmatrix}$$

Neumann series: $(\mathrm{Id} - A)^{-1} = \sum_{k=0}^{\infty} A^k$

- First Term: Simultaneous gradient descent; Optimal if other player stays still
- Second Term: Optimal if the other player thinks that the other player stays still
- Third Term: Optimal if the other player thinks that the other player thinks …

Why bilinear is the right notion of first order

The CGD update rule is similar to a regularized and damped Newton's method

$$\begin{pmatrix} x_{k+1} - x_k \\ y_{k+1} - y_k \end{pmatrix} = -\eta \begin{pmatrix} \mathrm{Id} + \eta D^2_{xx} f & \eta D^2_{xy} f \\ \eta D^2_{yx} g & \mathrm{Id} + D^2_{yy} g \end{pmatrix}^{-1} \begin{pmatrix} \nabla_x f \\ \nabla_y g \end{pmatrix}.$$

Why drop the diagonal block of the Hessian? We argue that **CGD is not a second order method, but the right way to do first order for competitive problems:**

Reason 1: Bilinear fully uses first order regularity

Most competitive optimization problems have objectives of the form

$$f(x,y) = \Phi\big(X(x), Y(y)\big),$$

for highly regular Φ possibly less regular $x \mapsto X(x)$, $y \mapsto Y(y)$ In this setting, mixed second derivatives are well behaved, as soon as gradients are well behaved.

Reason 2: It plays well with quadratic regularization

For non-convex f or g, Newton step can be local *worst* strategy of a player. This leads to spurious attractors of the dynamics. Bilinear approximation is highest order that always leads to local best strategies when using quadratic regularization.

Reason 3: It leads to the right invariance:

First or second order approximation are invariant under linear transformations:

$$(A^{-1}x)^T \nabla_x f(A\,\cdot) = x^T \nabla_x f, \qquad (A^{-1}x)^T D^2_{xx} f(A\,\cdot)\,(A^{-1}x) = x^T\, D^2_{xx} f\; x.$$

For the bilinear approximation only satisfies

$$\left(A^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right)^T \begin{pmatrix} 0 & D^2_{xy} f(A\,\cdot) \\ D^2_{yx} f(A\,\cdot) & 0 \end{pmatrix} \left(A^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 0 & D^2_{xy} f \\ D^2_{yx} f & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

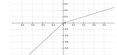
for $A = \begin{pmatrix} A_{xx} & 0 \\ 0 & A_{yy} \end{pmatrix}$ block diagonal. If A not block diagonal, it transfers decision variables between agents. This changes the game, don't want to be invariant to it!

Convergence results robust to strong interactions:

Consider zero sum game ($f = -g$) and define

$$\bar{D} := \left(\mathrm{Id} + \eta^2 D^2_{xy} f\, D^2_{yx} f\right)^{-1} \eta^2 D^2_{xy} f\, D^2_{yx} f \quad \bar{D} := \left(\mathrm{Id} + \eta^2 D^2_{xy} f\, D^2_{yx} f\right)^{-1} \eta^2 D^2_{yx} f D^2_{xy} f.$$

Define $h_\pm(\lambda) := \frac{\min(3\lambda, \lambda)}{2}$



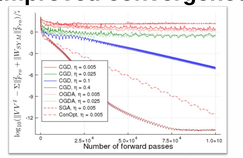
Theorem: If f is two times differentiable with L- Lipschitz continuous Hessian and $\eta||D^2_{xx} f||, \eta||D^2_{yy} f|| \leq 1$:

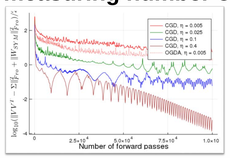
$$||\nabla_x f(x_{k+1}, y_{k+1})||^2 + ||\nabla_y f(x_{k+1}, y_{k+1})||^2 - ||\nabla_x f||^2 - ||\nabla_y f||^2$$
$$\leq -\nabla_x f^T\left(\eta h_\pm(D^2_{xx} f) + \bar{D} - 32L\eta^2||\nabla_x f||\right)\nabla_x f - \nabla_y f^T\left(\eta h_\pm(-D^2_{yy} f) + \bar{D} - 32L\eta^2||\nabla_y f||\right)\nabla_y f$$

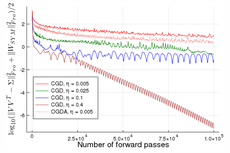
Strong interaction between the players only improves convergence!

Faster convergence and splitting of modes

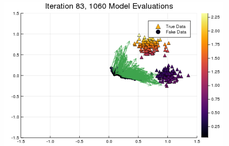
Improved convergence measuring number of model evaluations

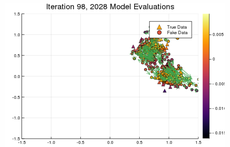






CGD rapidly splits mass onto different target modes, preventing mode collapse!





Replacing gradient penalty by CGD improves WGAN-GP Inception Score on CIFAR10: Come to the poster session *"Bridging Game Theory and Deep Learning", Sat Dec14 9:30 am – 11 am and 4:30 pm in West Exhibition Hall A!*