Omega: Optimistic EMA Gradients

Juan Ramirez, Rohan Sukumaran, Quentin Bertrand & Gauthier Gidel




Mila
Université de Montréal

Outline

- Stochastic Min-Max Optimization
    - The Optimistic Gradient Method
- Omega: Optimistic EMA Gradients
- Experiments
    - Stochastic *Bilinear* Games
    - Stochastic *Quadratic* Games
- Future Work

Min-Max Optimization

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \mathcal{L}(x, y) = \mathbb{E}_\xi \left[\ell(x, y, \xi) \right]$$

Min-Max Optimization

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \mathcal{L}(x, y) = \mathbb{E}_{\xi}\left[\ell(x, y, \xi)\right]$$

GANs

Adversarial robustness

Actor critic systems

Min-Max Optimization

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} \mathcal{L}(x, y) = \mathbb{E}_\xi \left[\ell(x, y, \xi) \right]$$

Goal: find a Nash equilibrium

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*)$$

Min-Max Optimization

Min-max optimization is generally harder than minimization!

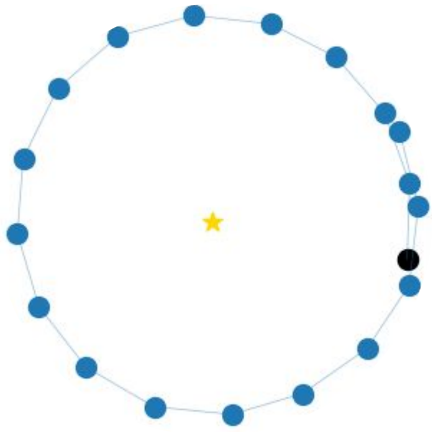
- Gradient descent - ascent (GDA) can diverge for convex - concave problems (Gidel et al., 2019).
- Presents oscillations.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In ICLR, 2019.

Min-Max Optimization

Min-max optimization is generally harder than minimization!

- Gradient descent - ascent (GDA) can diverge for convex - concave problems (Gidel et al., 2019).
- Presents oscillations.



G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In ICLR, 2019.

Min-Max Optimization

Min-max optimization is generally harder than minimization!

- Gradient descent - ascent (GDA) can diverge for convex - concave problems (Gidel et al., 2019).
- Presents oscillations.
- GDA and Extragradient are very sensitive to noise (Chavdarova et al., 2019).

T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. In NeurIPS, 2019.

Solving Stochastic Min-Max Problems

$$w = [x, y]^\top \quad F_\xi(w) = [\nabla_x \ell(x, y, \xi), -\nabla_y \ell(x, y, \xi)]^\top$$

Same-sample Extragradient

$$w_{t+1/2} = w_t - \eta F_{\xi_t}(w_t)$$
$$w_{t+1} = w_t - \eta F_{\xi_t}(w_{t+1/2})$$

Solving Stochastic Min-Max Problems

$$w = [x, y]^\top \quad F_\xi(w) = [\nabla_x \ell(x, y, \xi), -\nabla_y \ell(x, y, \xi)]^\top$$

Same-sample Extragradient

$$w_{t+1/2} = w_t - \eta F_{\xi_t}(w_t)$$
$$w_{t+1} = w_t - \eta F_{\xi_t}(w_{t+1/2})$$

Independent samples Optimistic Gradient

$$w_{t+1} = w_t - \eta \Big[F_{\xi_t}(w_t) + \alpha \big[F_{\xi_t}(w_t) - F_{\xi_{t-1}}(w_{t-1}) \big] \Big]$$

Store $F(w_t, \xi_t)$ for next update.

What About a *Same-Sample* Optimistic Gradient?

$$w_{t+1} = w_t - \eta\Big[F_{\xi_t}(w_t) + \alpha\big[F_{\xi_t}(w_t) - F_{\xi_t}(w_{t-1})\big]\Big]$$

What About a *Same-Sample* Optimistic Gradient?

$$w_{t+1} = w_t - \eta \Big[F_{\xi_t}(w_t) + \alpha \big[F_{\xi_t}(w_t) - F_{\xi_t}(w_{t-1}) \big] \Big]$$

- Can not re-use $F_{\xi_t}(w_t)$ in the next iteration, it will require $F_{\xi_{t+1}}(w_t)$
- "Same-Sample" Optimism requires **two gradients** per update, just like Extragradient!

Omega

$$w_{t+1} = w_t - \eta\Big[F_{\xi_t}(w_t) + \alpha[F_{\xi_t}(w_t) - \tilde{F}_{t-1}]\Big]$$

Where $\tilde{F}_t$ is an EMA of previously observed gradients:

$$\tilde{F}_t = (1-\beta)F_{\xi_t}(w_t) + \beta\tilde{F}_{t-1}$$

Intuitively, Omega should reduce the variance of the correction term, yielding a method that is robust to batch variances.

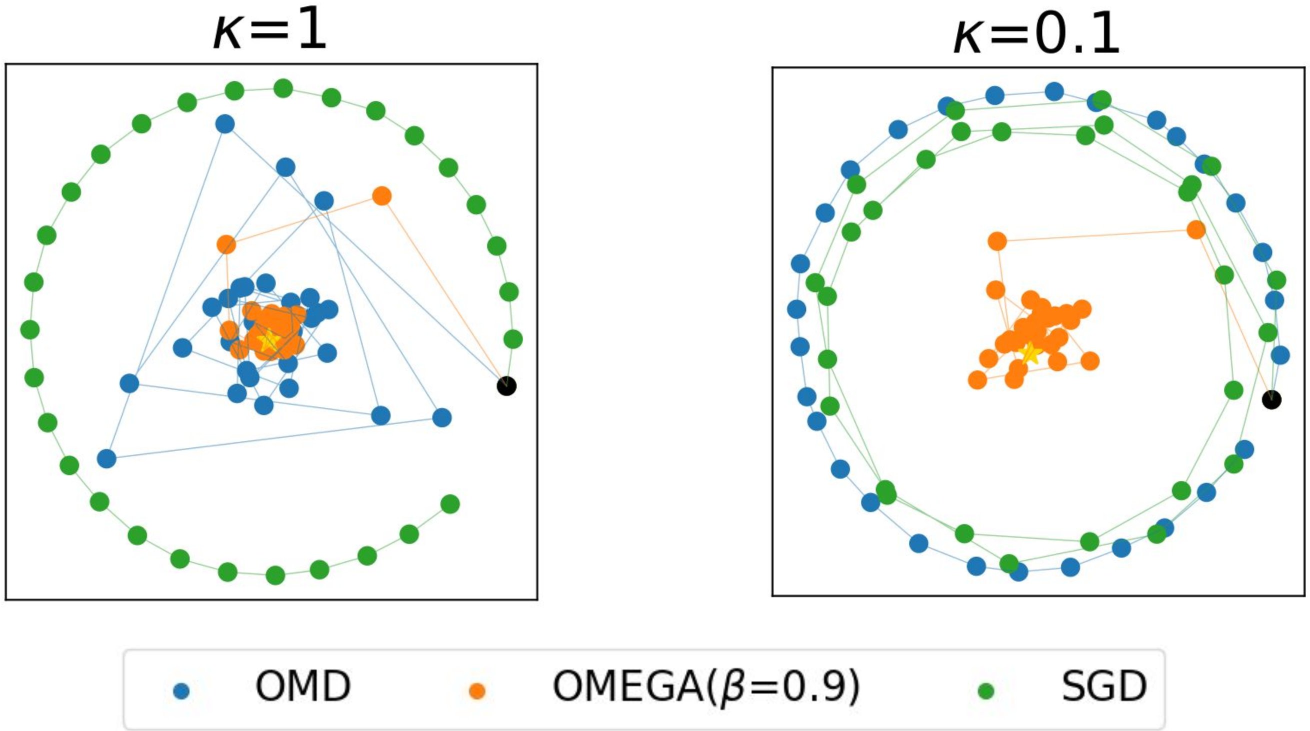
Stochastic Quadratic Games

$$\min_x \max_y \mathbb{E}_\xi \left[\frac{1}{2} x^\top A_\xi x + a_\xi x + x^\top B_\xi y - c_\xi y - \frac{1}{2} y^\top C_\xi y \right]$$

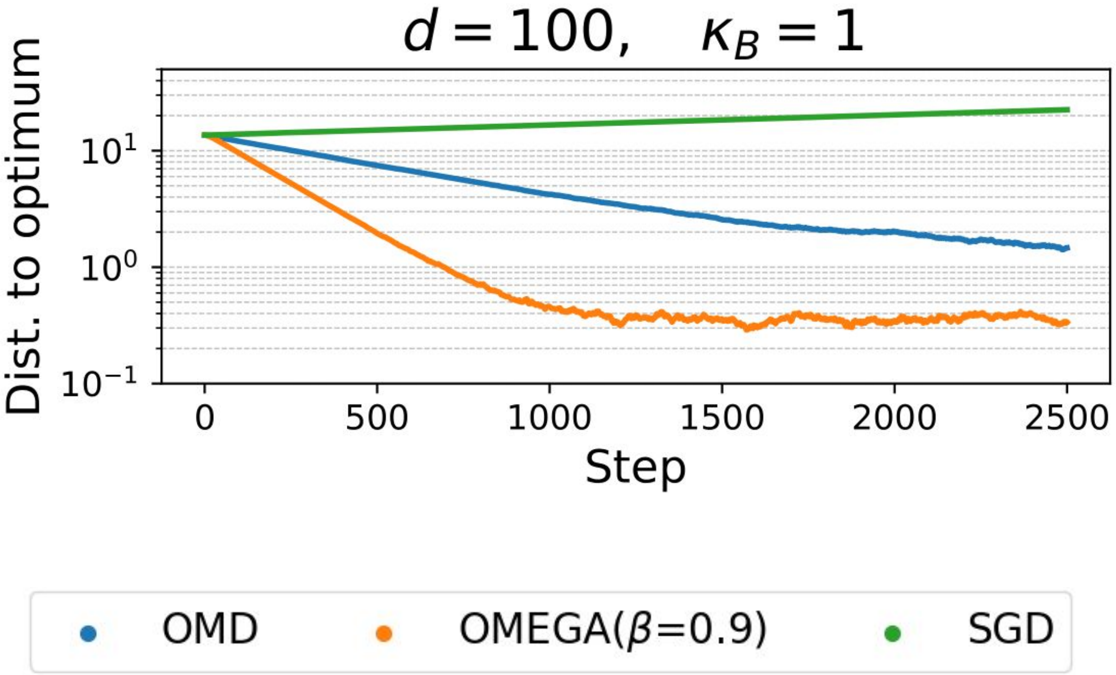
- Has a unique Nash equilibrium.
- Strongly convex - strongly concave for positive-definite A and C.

$$\mu_B I \preceq B \preceq L_B I \qquad \kappa_B = \frac{L_B}{\mu_B}$$

Bilinear

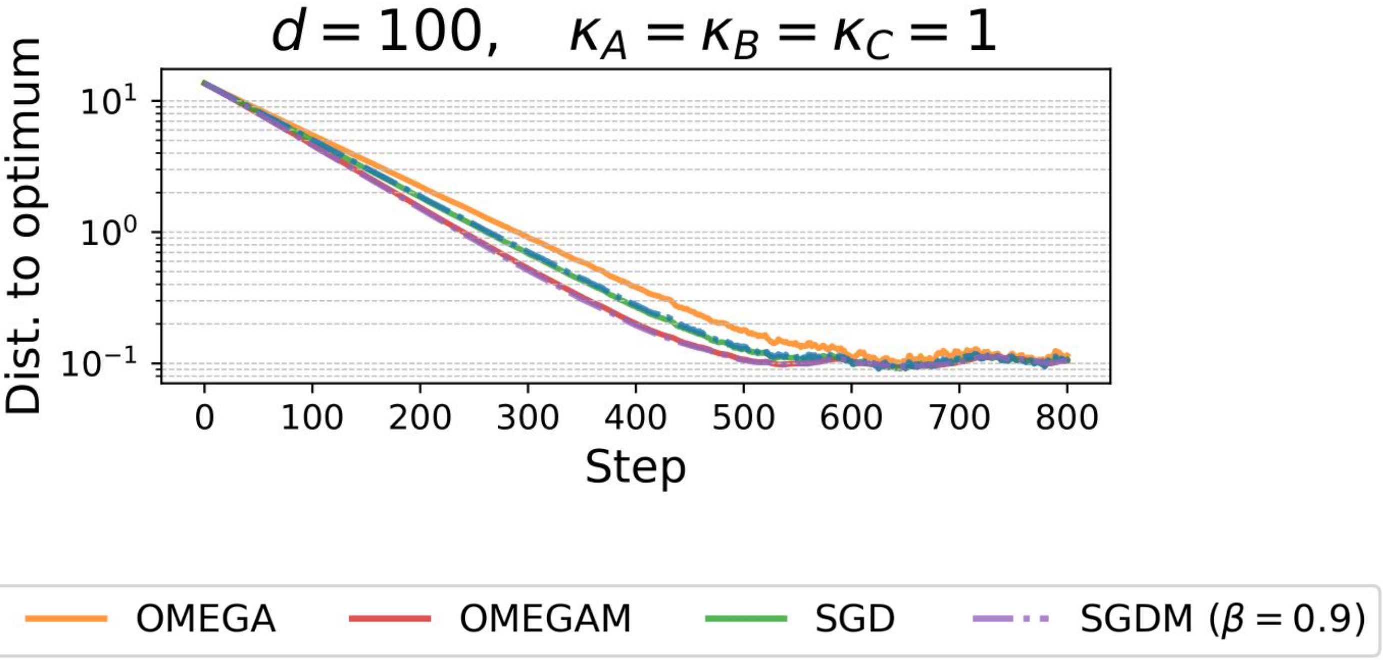


Bilinear



$d = 100, \quad \kappa_B = 1$
Dist. to optimum
10^1
10^0
10^{-1}
0
500
1000
1500
2000
2500
Step
OMD
OMEGA(β=0.9)
SGD

Quadratic



$d = 100, \quad \kappa_A = \kappa_B = \kappa_C = 1$
Dist. to optimum
10^1
10^0
10^{-1}
0
100
200
300
400
500
600
700
800
Step
OMD
OMEGA
OMEGAM
SGD
SGDM ($\beta = 0.9$)

Future Work

Omega shows promising results when applied to linear players.

- Policy Evaluation in RL (under a linear value function).
- Lagrangian-based constrained optimization.

Analyze the convergence properties of Omega.

References

- N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, & S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. In NeurIPS, 2021.
- L. D. Popov. A modification of the arrow-hurwicz method for search of saddle points. In Mathematical notes of the Academy of Sciences of the USSR, 1980.
- G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In ICLR, 2019.
- T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. In NeurIPS, 2019.