Momentum Extragradient is Optimal for Games with Cross-Shaped Spectrum

Junhyung Lyle Kim (Rice University), Gauthier Gidel (Mila & Université de Montréal), Anastasios Kyrillidis (Rice University), Fabian Pedregosa (Google Research)

[Differentiable Games]

- We focus on n-player differentiable game where the player i has the loss $\ell_i : \mathbb{R}^{d_i} \to \mathbb{R}$ for $i = 1, \ldots, n$ with the parameter $w^{(i)} \in \mathbb{R}^{d_i}$.
- We denote the concatenated parameters by $w = [w^{(1)}, \ldots, w^{(n)}] \in \mathbb{R}^d$, where $d = \sum_{i=1}^{n} d_i$.
- Vector field and Jacobian:

$$v(w) = [\nabla_{w^{(1)}} \ell_1(w), \ldots, \nabla_{w^{(n)}} \ell_n(w)]^\top$$

$$\nabla v(w) = \begin{bmatrix} \nabla^2_{w^{(1)}} \ell_1(w) & \ldots & \nabla_{w^{(n)}} \nabla_{w^{(1)}} \ell_1(w) \\ \vdots & & \vdots \\ \nabla_{w^{(1)}} \nabla_{w^{(n)}} \ell_n(w) & \ldots & \nabla^2_{w^{(n)}} \ell_1(w) \end{bmatrix}$$

- Goal: Find $w^\star \in \mathbb{R}^d$ such that $v(w^\star) = 0$.

[Residual Polynomials of EGM]

- Extragradient with Momentum (EGM):

$$w_{t+1} = w_t - hv(w_t - \gamma v(w_t)) + m(w_t - w_{t-1})$$

Theorem (Residual polynomials of EGM): Consider the EGM method. Its associated residual polynomials are:

$$\tilde{P}_0(\lambda) = 1, \quad \tilde{P}_1(\lambda) = 1 - \frac{h\lambda(1-\gamma\lambda)}{1+m}$$

$$\tilde{P}_{t+1}(\lambda) = (1 + m - h\lambda(1-\gamma\lambda))\tilde{P}_t(\lambda) - m\tilde{P}_{t-1}(\lambda).$$

Further, let $T_t(\cdot)$ and $U_t(\cdot)$ be the Chebyshev polynomials of the first and the second kind respectively. Then, the above expression simplifies to:

$$P_t(\lambda) = m^{t/2}\left(\frac{2m}{1+m}T_t(\sigma(\lambda)) + \frac{1-m}{1+m}U_t(\sigma(\lambda))\right)$$

with $\quad \sigma(\lambda) = \dfrac{1 + m - h\lambda(1-\gamma\lambda)}{2\sqrt{m}},$

where we refer to the term $\sigma(\lambda)$ as the link function.

Lemma: There exists a real polynomial p_t of degree at most t satisfying

$$w_t - w^\star = p_t(A)(w_0 - w^\star),$$

where $p_t(0) = 1$, and $v(w^\star) = Aw^\star + b$.

- Worst-case convergence rate:

$$\begin{aligned} \|w_t - w^\star\| &= \|p_t(A)(w_0 - w^\star)\| \\ &\le \|p_t(A)\| \cdot \|(w_0 - w^\star)\| \\ &\le \max_\lambda |p_t(\lambda)| \cdot \|(w_0 - w^\star)\| \end{aligned}$$

[Robust Region]

Lemma: Let z be any complex number. The sequence $\left(\left|\frac{2m}{1+m}T_t(z) + \frac{1-m}{1+m}U_t(z)\right|\right)_{t\ge0}$ grows exponentially in t for $z \notin [-1,1]$, while in that interval, they are bounded by $|T_t(z)| \le 1, \ |U_t(z)| \le t+1.$

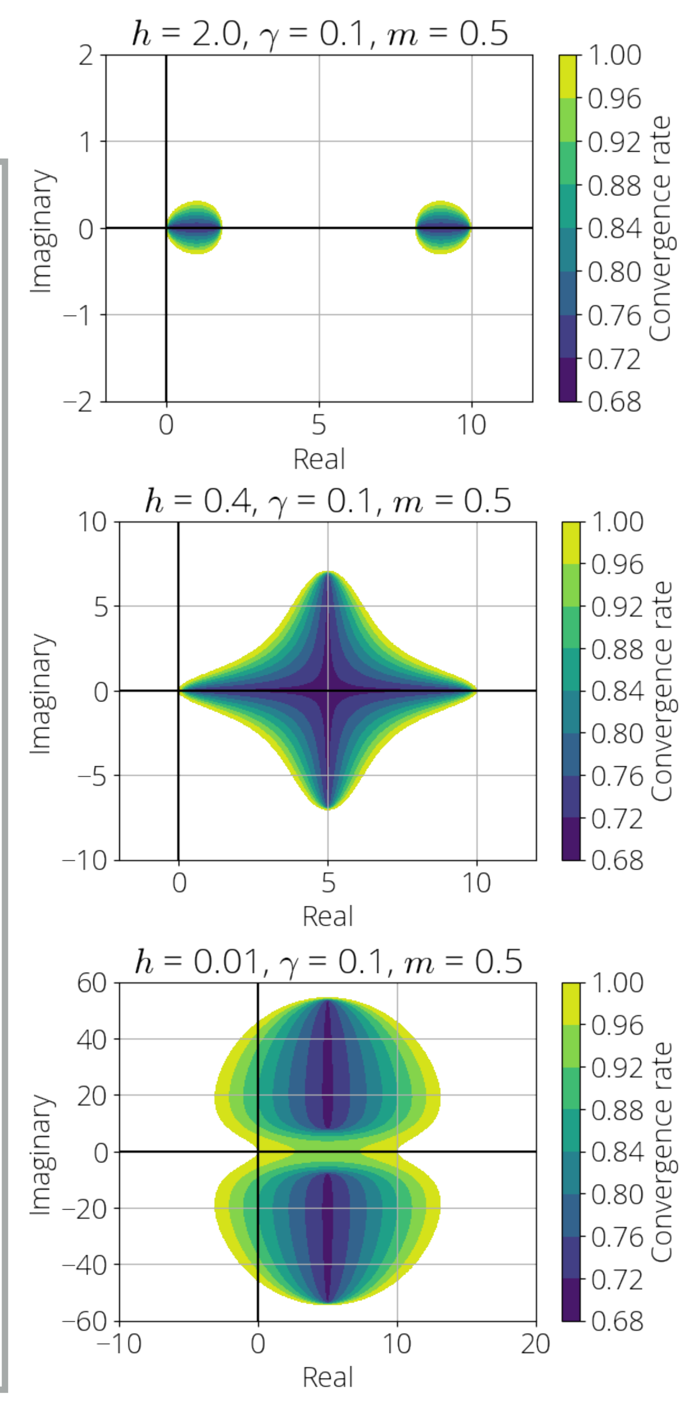
- Robust region: the set of hyperparameters where $\sigma(\lambda) \in [-1,1]$ $:= \sigma^{-1}([-1,1])$

$$\sigma^{-1}(1) = \frac{1}{2\gamma} \pm \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}} \qquad \sigma^{-1}(-1) = \frac{1}{2\gamma} \pm \sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}$$

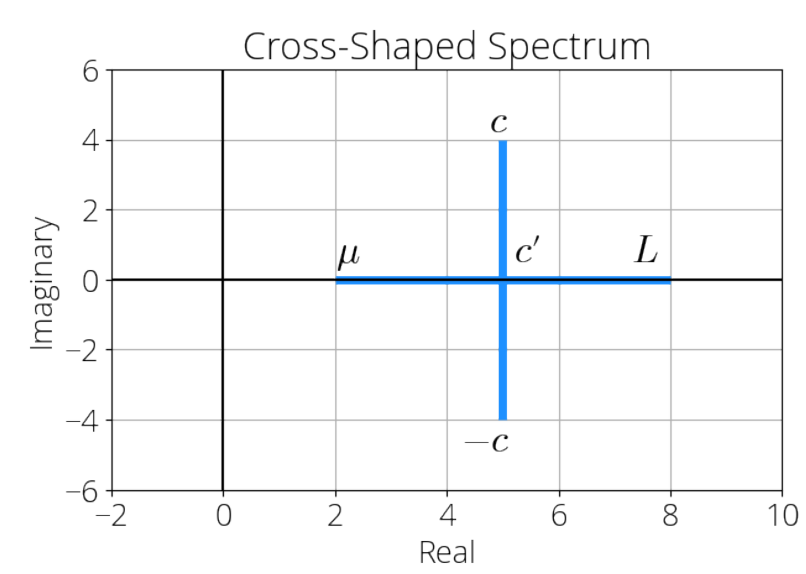
[Three Modes of EGM]

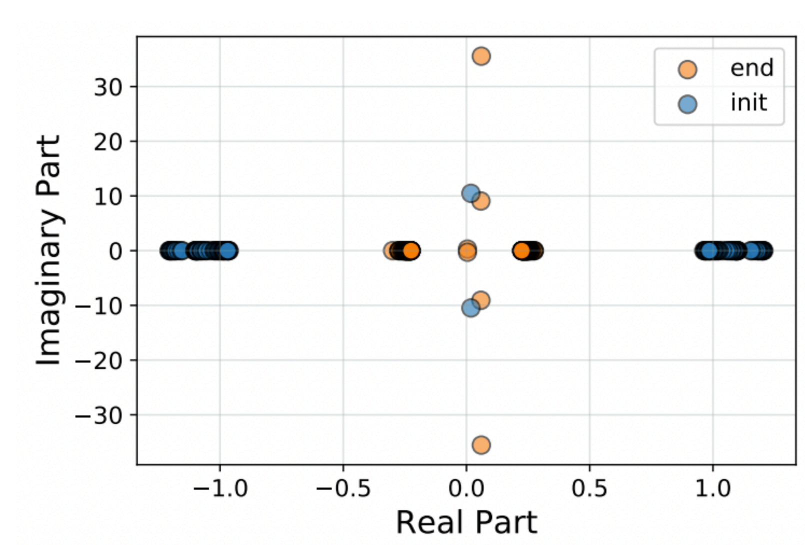
Theorem: Consider the EGM method. The robust region above have the following three modes:

1. If $\frac{h}{4\gamma} \ge \left(1+\sqrt{m}\right)^2$, then $\sigma^{-1}(-1)$ and $\sigma^{-1}(1)$ are all real numbers;
2. If $\left(1-\sqrt{m}\right)^2 \le \frac{h}{4\gamma} < \left(1+\sqrt{m}\right)^2$, then $\sigma^{-1}(-1)$ are complex, and $\sigma^{-1}(1)$ are real;
3. If $\left(1-\sqrt{m}\right)^2 > \frac{h}{4\gamma}$, then $\sigma^{-1}(-1)$ and $\sigma^{-1}(1)$ are all complex numbers.



[Games with Cross-Shaped Jacobian Spectrum]





$$\mathrm{Sp}(\nabla v(w)) \in \mathcal{S}^\star = [\mu, L] \cup \{a + bi \in \mathbb{C} : a = c' > 0, \ b \in [-c, c]\}$$

- Robust region:

$$\sigma^{-1}([-1,1]) = \left[\frac{1}{2\gamma} - \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}}, \ \frac{1}{2\gamma} + \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}}\right] \bigcup \left[\frac{1}{2\gamma} - \sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}, \ \frac{1}{2\gamma} + \sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}\right]$$

Theorem (Optimal hyperparameters for EGM): The optimal hyperparameters for EGM can be set as follows:

$$h = \frac{8(\mu+L)}{(\sqrt{\mu^2+L^2} + \sqrt{2\mu L})^2}, \quad \gamma = \frac{1}{\mu+L}, \quad \text{and} \quad m = \left(\frac{\sqrt{\mu^2+L^2} - \sqrt{2\mu L}}{\sqrt{\mu^2+L^2} + \sqrt{2\mu L}}\right)^2$$

- Worst-case asymptotic rate: $\limsup_{t\to\infty} \sqrt[2t]{r_t} = \sqrt[4]{m}$ where $r_t = \max_{\lambda \in \mathcal{S}^\star} |P_t(\lambda)|$.

$$\sqrt[4]{m} = \left(\frac{\sqrt{\mu^2+L^2} - \sqrt{2\mu L}}{\sqrt{\mu^2+L^2} + \sqrt{2\mu L}}\right)^{\frac{1}{2}} \underset{\tau\to0}{=} 1 - \sqrt{2}\sqrt{\tau} + o(\sqrt{\tau}) \approx 1 - \sqrt{2}\sqrt{\frac{\mu}{L}}.$$

- Gradient method and Extragradient method (without momentum) does not achieve an accelerated convergence rate