Dealing With Unbounded Gradients in Stochastic Saddle-point Optimization

Gergely Neu, Nneka Okolo

July 17, 2024

Universitat Pompeu Fabra Barcelona

ICML International Conference On Machine Learning

Bilinear game:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y}) := \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{y}\rangle_{\mathbb{R}^m} + \langle \boldsymbol{b}, \boldsymbol{x}\rangle_{\mathbb{R}^m} - \langle \boldsymbol{c}, \boldsymbol{y}\rangle_{\mathbb{R}^n} . \tag{1}$$

Goal:

- Find a saddle-point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ such that,

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \le f(\boldsymbol{x}^*, \boldsymbol{y}^*) \le f(\boldsymbol{x}, \boldsymbol{y}^*). \tag{2}$$

Stochastic Gradient Descent-Ascent (SGDA)

Compute:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_x \widetilde{\boldsymbol{g}}_x(t),$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_y \widetilde{\boldsymbol{g}}_y(t).$$

But,

$$\widetilde{\boldsymbol{g}}_x(t) = (\boldsymbol{M} + \xi_{\boldsymbol{M}}(t))\, \boldsymbol{y}_t + (\boldsymbol{b} + \xi_b(t)),$$

and

$$\widetilde{\boldsymbol{g}}_y(t) = (\boldsymbol{M} + \xi_M(t))^{\top} \boldsymbol{x}_t - (\boldsymbol{c} + \xi_c(t)).$$

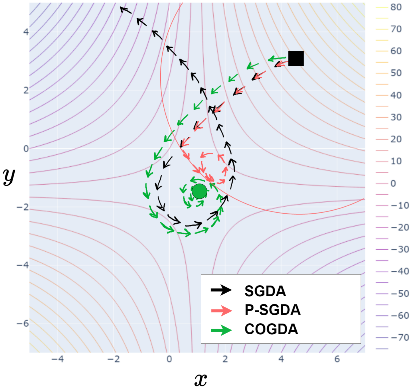

Figure: Illustration of SGDA, P-SGDA and COGDA on an example Bilinear game with $f(x, y) = (x - 1) * (y + 1.5)$.

SGDA with Projections (P-SGDA)

Compute:

$$x_{t+1} = P_{\mathbb{B}(D_{\mathcal{X}})}\left(x_t - \eta_x \widetilde{g}_x(t)\right),$$
$$y_{t+1} = P_{\mathbb{B}(D_{\mathcal{Y}})}\left(y_t + \eta_y \widetilde{g}_y(t)\right).$$

But, we need further knowledge of (x^*, y^*) to properly set $D_{\mathcal{X}}, D_{\mathcal{Y}}$.

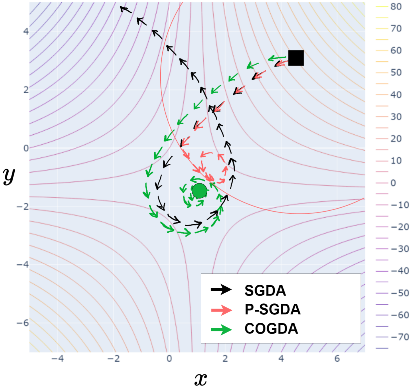

Figure: Illustration of SGDA, P-SGDA and COGDA on an example Bilinear game with $f(x, y) = (x - 1) * (y + 1.5)$.

Composite Objective Gradient Descent Ascent (COGDA)

Compute:

$$x_{t+1} = \frac{x_t - \eta_x \widetilde{g}_x(t)}{1 + \varrho_x \eta_x} + \frac{\varrho_x \eta_x x_1}{1 + \varrho_x \eta_x},$$

$$y_{t+1} = \frac{y_t + \eta_y \widetilde{g}_y(t)}{1 + \varrho_y \eta_y} + \frac{\varrho_y \eta_y y_1}{1 + \varrho_y \eta_y}.$$

No need for standard assumptions such as:

- ✗ Prior knowledge of $\|x^*\|_2$ (resp. $\|y^*\|_2$),
- ✗ f is G-Lipschitz (with known G),
- ✗ Noise is uniformly bounded or **light-tailed**.

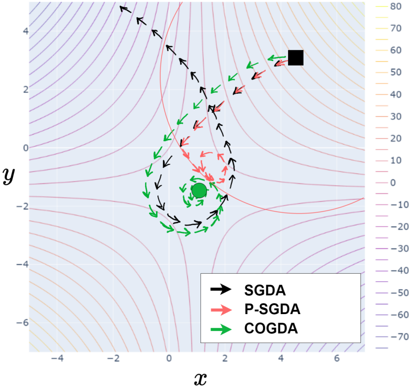

Figure: Illustration of SGDA, P-SGDA and COGDA on an example Bilinear game with $f(x, y) = (x - 1) * (y + 1.5)$.

For $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}$,

$$
\begin{aligned}
&\mathbb{E}\left[G(\boldsymbol{x}^*, \boldsymbol{y}^*)\right] \\
&\quad \leq \frac{\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2}{2\eta_x T} + \frac{\eta_x}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}_x(t)\|_2^2\right] + \frac{\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2}{2\eta_y T} + \frac{\eta_y}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}_y(t)\|_2^2\right] \\
&\qquad + \frac{\varrho_x}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2 - \|\boldsymbol{x}_t - \boldsymbol{x}_1\|_2^2\right] + \frac{\varrho_y}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_1\|_2^2\right].
\end{aligned}
$$

Convergence guarantee for COGDA

After at most T iterations, with $\boldsymbol{x}_1, \boldsymbol{y}_1 = 0$ as well as $\eta_y, \eta_x = 1/D_M\sqrt{T}$ and $\varrho_y, \varrho_x = 2D_M/\sqrt{T}$. Then,

$$
\mathbb{E}\left[G(\boldsymbol{x}^*, \boldsymbol{y}^*)\right] = \mathcal{O}\left(\frac{\|\boldsymbol{x}^*\|_2^2 + \|\boldsymbol{y}^*\|_2^2 + 1}{\sqrt{T}}\right).
$$

More details in the paper

- ✓ Our guarantees hold for data-dependent comparators $(\boldsymbol{x}^*, \boldsymbol{y}^*)$,
- ✓ An approach for Sub-bilinear functions f - which behaves like a bilinear function asymptotically as one approaches infinity in each axis,
- ✓ Application to planning in tabular Average-reward Markov Decision Processes without prior knowledge of the bias span.

Thank You