Parameter-free (Second-order) Methods for Min-max Optimization

Ali Kavis

UT Austin

Joint work with:
Ruichen Jiang, Qiujiang Jin, Sujay Sanghavi, Aryan Mokhtari

General problem setting

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y})$$

General problem setting

$$\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y})$$

Examples

Primal-dual optimization

GANs

Reinforcement learning

Multi-agent games

Examples

Convex optimization with equality constraints

$$\min_{\mathbf{x}} \; f(\mathbf{x})$$
$$\text{s.t. } \; \mathbf{A}\mathbf{x} = \mathbf{b}$$

$$\Downarrow$$

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle$$

Examples

Generative adversarial networks (GANs)

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_{\text{noise}}}[1 - \log(D(G(\mathbf{z})))]$$

Karras et al., PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION. ICLR, 2018.

Optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y})$$

Optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y})$$

$f(\cdot, \mathbf{y})$ convex in **x** for all **y**

$f(\mathbf{x}, \cdot)$ concave in **y** for all **x**

Restricted gap

$$\mathrm{Gap}_{\mathcal{X}\times\mathcal{Y}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) := \max_{\mathbf{y}\in\mathcal{X}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{Y}} f(\mathbf{x}, \hat{\mathbf{y}})$$

Optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y})$$

$f(\cdot,\mathbf{y})$ convex in **x** for all **y**

$f(\mathbf{x},\cdot)$ concave in **y** for all **x**

Restricted gap

$$\mathrm{Gap}_{\mathcal{X}\times\mathcal{Y}}(\hat{\mathbf{x}},\hat{\mathbf{y}}) := \max_{\mathbf{y}\in\mathcal{X}} f(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{Y}} f(\mathbf{x},\hat{\mathbf{y}})$$

$$f(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y}\rangle$$

Operator Representation

$$\mathbf{z} = (\mathbf{x}, \mathbf{y}) \qquad\qquad \mathbf{F}(\mathbf{z}) := \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix}$$

Operator Representation

$$\mathbf{z} = (\mathbf{x}, \mathbf{y}) \qquad \mathbf{F}(\mathbf{z}) := \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix}$$

f is convex-concave <—> F is monotone

$$\langle \mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2), \mathbf{z}_1 - \mathbf{z}_2 \rangle \geq 0$$

New formulation

$$\text{Find } \mathbf{z}^* \text{ s.t.} \quad \langle \mathbf{F}(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \quad \forall \mathbf{z} \in \mathbb{R}^{m+n}$$

How do we quantify performance?

$$\text{Generate} \quad \{\mathbf{z}_t\}_{t=1}^T \qquad \text{—>} \qquad \text{Compute Regret} := \sum_{t=1}^T \theta_t \langle \mathbf{F}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z} \rangle$$

$$f(\bar{\mathbf{x}}_T, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_T) \leq \sum_{t=1}^T \theta_t \langle \mathbf{F}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z} \rangle$$

Recap: **Lipschitz gradient** setting

f has L1-Lipschitz gradient <—> F is L1-Lipschitz

$$\|\mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2)\| \leq L_1 \|\mathbf{z}_1 - \mathbf{z}_2\|$$

Lower bound [Nemirovski, 1992]**:**

When **F** is monotone and Lipschitz continuous **—>** $\Omega\left(\frac{1}{T}\right)$

Recap: **Lipschitz gradient** setting

f has L1-Lipschitz gradient <—> F is L1-Lipschitz

$$\|\mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2)\| \leq L_1 \|\mathbf{z}_1 - \mathbf{z}_2\|$$

Lower bound [Nemirovski, 1992]**:**

When **F** is monotone and Lipschitz continuous **—>** $\Omega\left(\frac{1}{T}\right)$

Many algorithms have matching upper complexity bounds

- Extragradient [Korpelevich, 1978] **&** Mirror-Prox [Nemirovski, 2004]

$$\mathbf{z}_{t+1/2} = \mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_t)$$
$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_{t+1/2})$$

- Optimistic Method **&** Forward-Reflected-Backward [Malitsky & Tam, 2020]

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \left[\eta_t \mathbf{F}(\mathbf{z}_t) + \eta_{t-1}\left(\mathbf{F}(\mathbf{z}_t) - \mathbf{F}(\mathbf{z}_{t-1})\right)\right]$$

Our focus: **Lipschitz Hessian** setting

f has L2-Lipschitz Hessian <—> F' is L2-Lipschitz

$$\|\mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2) - \mathbf{F}'(\mathbf{z}_2)(\mathbf{z}_1 - \mathbf{z}_2)\| \leq \frac{L_2}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2$$

Lower bound [Adil et al., 2022]**:**

When **F** is monotone and **F'** is Lipschitz continuous **—>** $\Omega\left(\frac{1}{T^{1.5}}\right)$

Our focus: **Lipschitz Hessian** setting

f has L2-Lipschitz Hessian <—> F' is L2-Lipschitz

$$\|\mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2) - \mathbf{F}'(\mathbf{z}_2)(\mathbf{z}_1 - \mathbf{z}_2)\| \leq \frac{L_2}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2$$

Lower bound [Adil et al., 2022]**:**

When **F** is monotone and **F'** is Lipschitz continuous **—>** $\Omega\left(\frac{1}{T^{1.5}}\right)$

2 classes of second-order algorithms

- **Line search methods —>** *need to take large enough steps*

  Iterative line search subroutine **—> extra computation**

- **Sub solver-based methods —>** *iterate update does not have a closed form*

  Iterative sub solver to compute the next iterate **—> extra computation**

Our focus: **Lipschitz Hessian** setting

f has L2-Lipschitz Hessian <—> F' is L2-Lipschitz

$$\|\mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2) - \mathbf{F}'(\mathbf{z}_2)(\mathbf{z}_1 - \mathbf{z}_2)\| \leq \frac{L_2}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2$$

Lower bound [Adil et al., 2022]**:**

When **F** is monotone and **F'** is Lipschitz continuous **—>** $\Omega\left(\frac{1}{T^{1.5}}\right)$

2 classes of second-order algorithms

- **Line search methods —>** *need to take large enough steps*

  Iterative line search subroutine **—> extra computation**

- **Sub solver-based methods —>** *iterate update does not have a closed form*

  Iterative sub solver to compute the next iterate **—> extra computation**

Our focus: **Lipschitz Hessian** setting

f has L2-Lipschitz Hessian <—> F' is L2-Lipschitz

$$\|\mathbf{F}(\mathbf{z}_1) - \mathbf{F}(\mathbf{z}_2) - \mathbf{F}'(\mathbf{z}_2)(\mathbf{z}_1 - \mathbf{z}_2)\| \leq \frac{L_2}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2$$

Lower bound [Adil et al., 2022]**:**

When **F** is monotone and **F'** is Lipschitz continuous **—>** $\Omega\left(\frac{1}{T^{1.5}}\right)$

Our Goal

- Achieve **optimal rate**
- **No line search** and **no sub solver** required
- **No need** to know problem **parameters**
- **Simple** and **intuitive** algorithm

Optimistic method and Proximal point method

Proximal point method (PPM) is an implicit method

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_{t+1})$$

Optimistic method and Proximal point method

Proximal point method (PPM) is an implicit method

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t \mathbf{F}(\mathbf{z}_{t+1})$$

How does optimism work?

Predict $\mathbf{F}(\mathbf{z}_{t+1})$ using the current information at $\mathbf{z}_t$

Correct the prediction by using the *previous* prediction terms at time $t-1$

Scale the prediction and correction terms with η_t and η_{t-1}, respectively

Second-order optimistic method

Prediction term is a linearization by means of the second-order term

$$\mathbf{F}(\mathbf{z}_{t+1}) \approx (\mathbf{F}(\mathbf{z}_t) + \mathbf{F}'(\mathbf{z}_t)(\mathbf{z}_{t+1} - \mathbf{z}_t))$$

Correction = Current term - Previous prediction

$$\mathbf{e}_t := \mathbf{F}(\mathbf{z}_t) - \mathbf{F}(\mathbf{z}_{t-1}) - \mathbf{F}'(\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{z}_{t-1})$$

Combine the prediction and correction terms

$$\eta_t \mathbf{F}(\mathbf{z}_{t+1}) \approx \eta_t [\mathbf{F}(\mathbf{z}_t) + \mathbf{F}'(\mathbf{z}_t)(\mathbf{z}_{t+1} - \mathbf{z}_t)] + \eta_{t-1}\mathbf{e}_t$$

Rearranging the terms will yield the update rule

$$\mathbf{z}_{t+1} = \mathbf{z}_t - (\mathbf{I} + \eta_t \mathbf{F}'(\mathbf{z}_t))^{-1}(\eta_t \mathbf{F}(\mathbf{z}_t) + \eta_{t-1}\mathbf{e}_t)$$

Why linesearch?

Guided by the analysis, we need to satisfy an error condition:

$$\eta_t \|\mathbf{e}_{t+1}\| \leq \alpha \|\mathbf{z}_{t+1} - \mathbf{z}_t\|$$

PPM satisfies the condition with $\alpha = 0$

Optimistic method requires $\alpha \leq 0.5$

Why linesearch?

Guided by the analysis, we need to satisfy an error condition:

$$\eta_t \|\mathbf{e}_{t+1}\| \leq \alpha \|\mathbf{z}_{t+1} - \mathbf{z}_t\|$$

PPM satisfies the condition with $\alpha = 0$

Optimistic method requires $\alpha \leq 0.5$

However, the condition itself is implicit

We can compute $\mathbf{z}_{t+1}$ **only after** we commit to η_t

Naive choice: **small** η_t **—>** **slow** progress and rate

We need η_t to be **large**, but still **satisfy** the error **condition**

Our goal

Lower bound [Adil et al., 2022]**:**

When **F** is monotone and **F'** is Lipschitz continuous **—>** $\Omega\left(\frac{1}{T^{1.5}}\right)$

Our Goal

- Achieve **optimal rate** (large enough step size)
- **No line search** and **no sub solver** required
- **No need** to know problem **parameters** (parameter-independent step size)
- **Simple** and **intuitive** algorithm

The algorithm

Algorithm 1 Adaptive Second-order Optimistic Method

1: **Input**: Initial points $\mathbf{z}_0 = \mathbf{z}_1 \in \mathbb{R}^m \times \mathbb{R}^n$, initial parameters $\eta_0 = 0$ and $\lambda_0 > 0$
2: **for** $t = 1, \ldots, T$ **do**
3: Set: $\mathbf{e}_t = \mathbf{F}(\mathbf{z}_t) - \mathbf{F}(\mathbf{z}_{t-1}) - \mathbf{F}'(\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{z}_{t-1})$
4: Set the step size parameters

$$\lambda_t = \begin{cases} L_2 & \textbf{(I)} \\ \max\left\{\lambda_{t-1}, \frac{2\|\mathbf{e}_t\|}{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}\right\} & \textbf{(II)} \end{cases} \qquad \eta_t = \frac{\lambda_t}{2\left(\eta_{t-1}\|\mathbf{e}_t\| + \sqrt{\eta_{t-1}^2\|\mathbf{e}_t\|^2 + \lambda_t\|\mathbf{F}(\mathbf{z}_t)\|}\right)}$$

5: Update: $\mathbf{z}_{t+1} = \mathbf{z}_t - (\lambda_t \mathbf{I} + \eta_t \mathbf{F}'(\mathbf{z}_t))^{-1} (\eta_t \mathbf{F}(\mathbf{z}_t) + \eta_{t-1}\mathbf{e}_t)$
6: **end for**
7: **return** $\bar{\mathbf{z}}_{T+1} = (\sum_{t=0}^{T} \eta_t)^{-1} \sum_{t=0}^{T} \eta_t \mathbf{z}_{t+1}$

The algorithm

Algorithm 1 Adaptive Second-order Optimistic Method

1: **Input**: Initial points $\mathbf{z}_0 = \mathbf{z}_1 \in \mathbb{R}^m \times \mathbb{R}^n$, initial parameters $\eta_0 = 0$ and $\lambda_0 > 0$

2: **for** $t = 1, \ldots, T$ **do**

3: Set: $\mathbf{e}_t = \mathbf{F}(\mathbf{z}_t) - \mathbf{F}(\mathbf{z}_{t-1}) - \mathbf{F}'(\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{z}_{t-1})$

4: Set the step size parameters

$$\lambda_t = \begin{cases} L_2 & \textbf{(I)} \\ \max\left\{\lambda_{t-1}, \frac{2\|\mathbf{e}_t\|}{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}\right\} & \textbf{(II)} \end{cases} \qquad \eta_t = \frac{\lambda_t}{2\left(\eta_{t-1}\|\mathbf{e}_t\| + \sqrt{\eta_{t-1}^2\|\mathbf{e}_t\|^2 + \lambda_t\|\mathbf{F}(\mathbf{z}_t)\|}\right)}$$

5: Update: $\mathbf{z}_{t+1} = \mathbf{z}_t - (\lambda_t \mathbf{I} + \eta_t \mathbf{F}'(\mathbf{z}_t))^{-1} (\eta_t \mathbf{F}(\mathbf{z}_t) + \eta_{t-1}\mathbf{e}_t)$

6: **end for**

7: **return** $\bar{\mathbf{z}}_{T+1} = (\sum_{t=0}^{T} \eta_t)^{-1} \sum_{t=0}^{T} \eta_t \mathbf{z}_{t+1}$

(Scaled) error condition - constant step size

We compute the regret and simplify the expression:

$$\sum_{t=1}^{T} \eta_t \langle \mathbf{F}(\mathbf{z}_{t+1}), \mathbf{z}_{t+1} - \mathbf{z} \rangle \leq \frac{\lambda}{2} \|\mathbf{z}_1 - \mathbf{z}\|^2 - \frac{\lambda}{4} \|\mathbf{z}_{T+1} - \mathbf{z}\|^2$$

$$+ \sum_{t=1}^{T} \frac{\eta_t^2}{\lambda} \|\mathbf{e}_{t+1}\|^2 - \frac{\lambda}{4} \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$$

For small enough α , we want to achieve:

$$\frac{\eta_t}{\lambda} \|\mathbf{e}_{t+1}\| \leq \alpha \|\mathbf{z}_{t+1} - \mathbf{z}_t\|$$

(Scaled) error condition - constant step size

Rearrange the terms in the error condition

$$\frac{\eta_t}{\lambda}\|\mathbf{e}_{t+1}\| \leq \alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \quad \longleftrightarrow \quad \frac{\eta_t\|\mathbf{e}_{t+1}\|}{\lambda\alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\|} \leq 1$$

(Scaled) error condition - constant step size

Rearrange the terms in the error condition

$$\frac{\eta_t}{\lambda}\|\mathbf{e}_{t+1}\| \leq \alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \quad \longleftrightarrow \quad \frac{\eta_t\|\mathbf{e}_{t+1}\|}{\lambda\alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\|} \leq 1$$

By the smoothness inequality and the update rule

$$\|\mathbf{e}_{t+1}\| = \|\mathbf{F}(\mathbf{z}_{t+1}) - \mathbf{F}(\mathbf{z}_t) - \mathbf{F}'(\mathbf{z}_t)(\mathbf{z}_{t+1} - \mathbf{z}_t)\| \leq \frac{L_2}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$$

$$\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \leq \frac{1}{\lambda_t}\eta_t\|\mathbf{F}(\mathbf{z}_t)\| + \frac{1}{\lambda_t}\eta_{t-1}\|\mathbf{e}_t\|$$

(Scaled) error condition - constant step size

Rearrange the terms in the error condition

$$\frac{\eta_t}{\lambda}\|\mathbf{e}_{t+1}\| \leq \alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \quad \longleftrightarrow \quad \frac{\eta_t\|\mathbf{e}_{t+1}\|}{\lambda\alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\|} \leq 1$$

By the smoothness inequality and the update rule

$$\|\mathbf{e}_{t+1}\| = \|\mathbf{F}(\mathbf{z}_{t+1}) - \mathbf{F}(\mathbf{z}_t) - \mathbf{F}'(\mathbf{z}_t)(\mathbf{z}_{t+1} - \mathbf{z}_t)\| \leq \frac{L_2}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$$

$$\|\mathbf{z}_{t+1} - \mathbf{z}_t\| \leq \frac{1}{\lambda_t}\eta_t\|\mathbf{F}(\mathbf{z}_t)\| + \frac{1}{\lambda_t}\eta_{t-1}\|\mathbf{e}_t\|$$

Combining all implies the following inequality

$$\frac{\eta_t\|\mathbf{e}_{t+1}\|}{\lambda\alpha\|\mathbf{z}_{t+1} - \mathbf{z}_t\|} \leq \frac{L_2\eta_t(\eta_t\|\mathbf{F}(\mathbf{z}_t)\| + \eta_{t-1}\|\mathbf{e}_t\|)}{2\alpha\lambda^2} = 1 \quad \longleftarrow$$

Explicit
Quadratic in η_t

Convergence theorem (constant step size)

Assumptions

- **F** is monotone (convex-concave)
- **F'** is L_2-Lipschitz (Hessian Lipschitz)
- **Option I:** $\lambda_t = L_2$

Convergence results
Bounded iterates
$$\|\mathbf{z}_t - \mathbf{z}^*\| \leq \frac{2}{\sqrt{3}}\|\mathbf{z}_1 - \mathbf{z}^*\|$$
Convergence rate
$$\mathrm{Gap}_{\mathcal{X}\times\mathcal{Y}}(\bar{\mathbf{z}}_{T+1}) \leq O\left(\frac{L_2\|\mathbf{z}_0 - \mathbf{z}^*\| \sup_{\mathbf{z}\in\mathcal{X}\times\mathcal{Y}} \|\mathbf{z}_1 - \mathbf{z}\|^2}{T^{1.5}}\right)$$

Convergence theorem (parameter-free)

Assumptions

- **F** is monotone (convex-concave)
- **F** is L_1-Lipschitz (gradient Lipschitz)
- **F'** is L_2-Lipschitz (Hessian Lipschitz)
- **Option II:** $\lambda_t = \max\left\{\lambda_{t-1}, \frac{2\|\mathbf{e}_t\|}{\|\mathbf{z}_{t-1} - \mathbf{z}_t\|^2}\right\}$

Convergence results

Bounded iterates

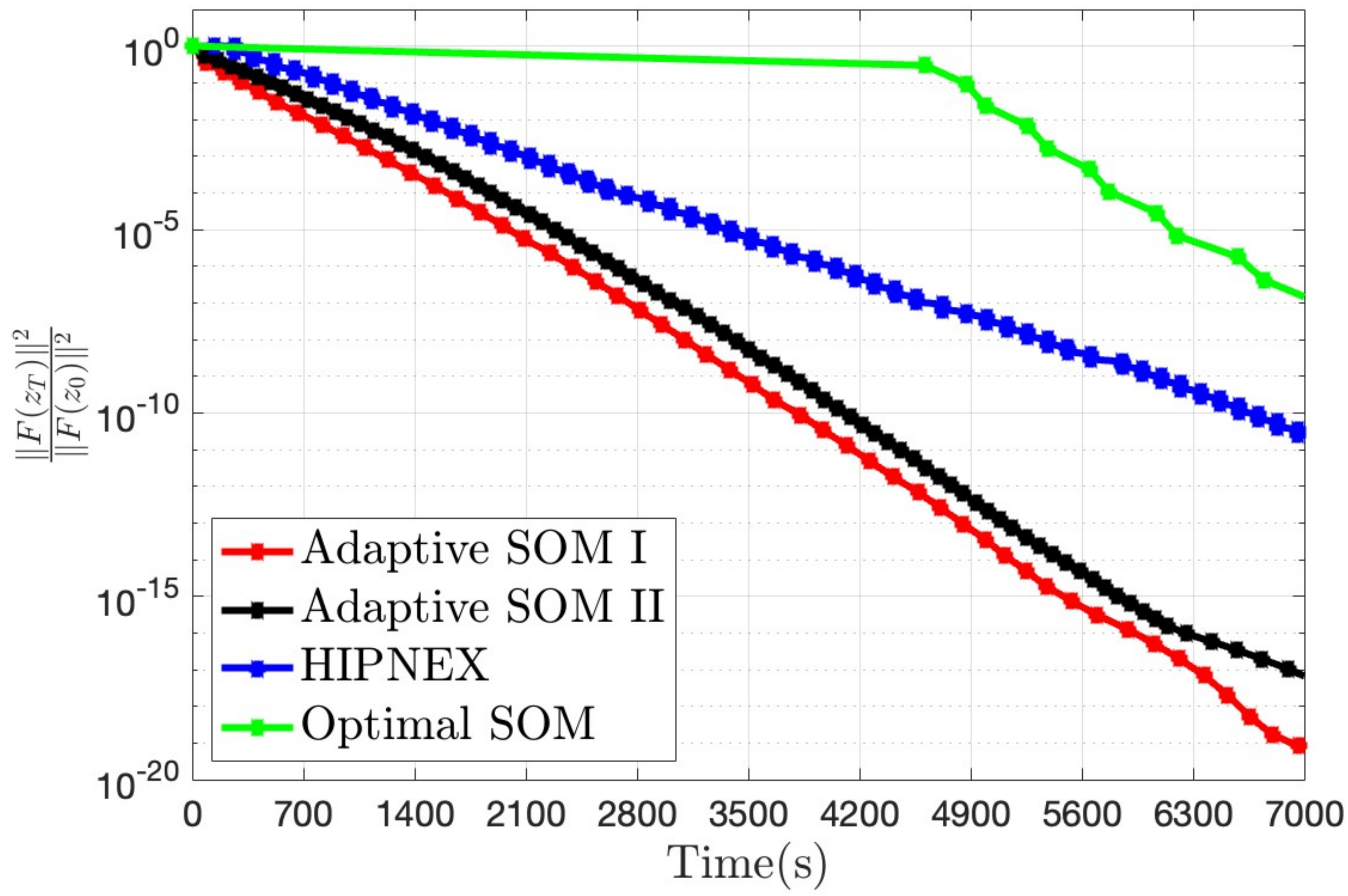
$$\|\mathbf{z}_t - \mathbf{z}^*\| \leq D \quad \text{where} \quad D^2 = \frac{L_1^2}{\lambda_1^2} + 2\frac{L_2^2}{\lambda_1^2}\|\mathbf{z}_1 - \mathbf{z}^*\|^2$$

Convergence rate

$$\mathrm{Gap}_{\mathcal{X}\times\mathcal{Y}}(\bar{\mathbf{z}}_{T+1}) \leq O\left(\frac{L_2\|\mathbf{z}_1 - \mathbf{z}^*\|\left(\sup_{\mathbf{z}\in\mathcal{X}\times\mathcal{Y}}\|\mathbf{z} - \mathbf{z}^*\|^2 + \frac{5}{4}D^2\right)}{T^{1.5}}\right)$$

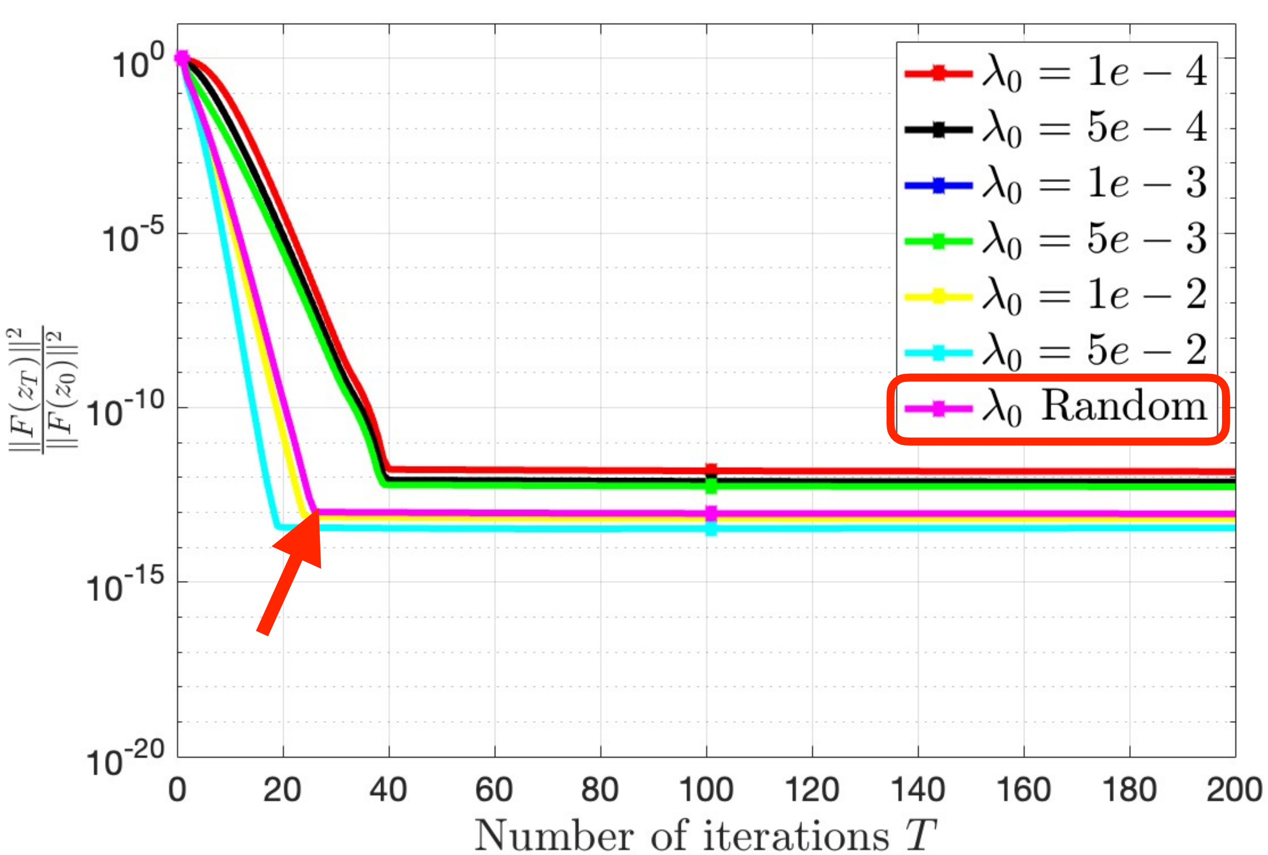
Some experiments

$$\min_{\mathbf{x}\in\mathbb{R}^n}\max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y}) = (\mathbf{A}\mathbf{x}-\mathbf{b})^\top\mathbf{y} + (L_2/6)\|\mathbf{x}\|^{\mathbf{3}}$$

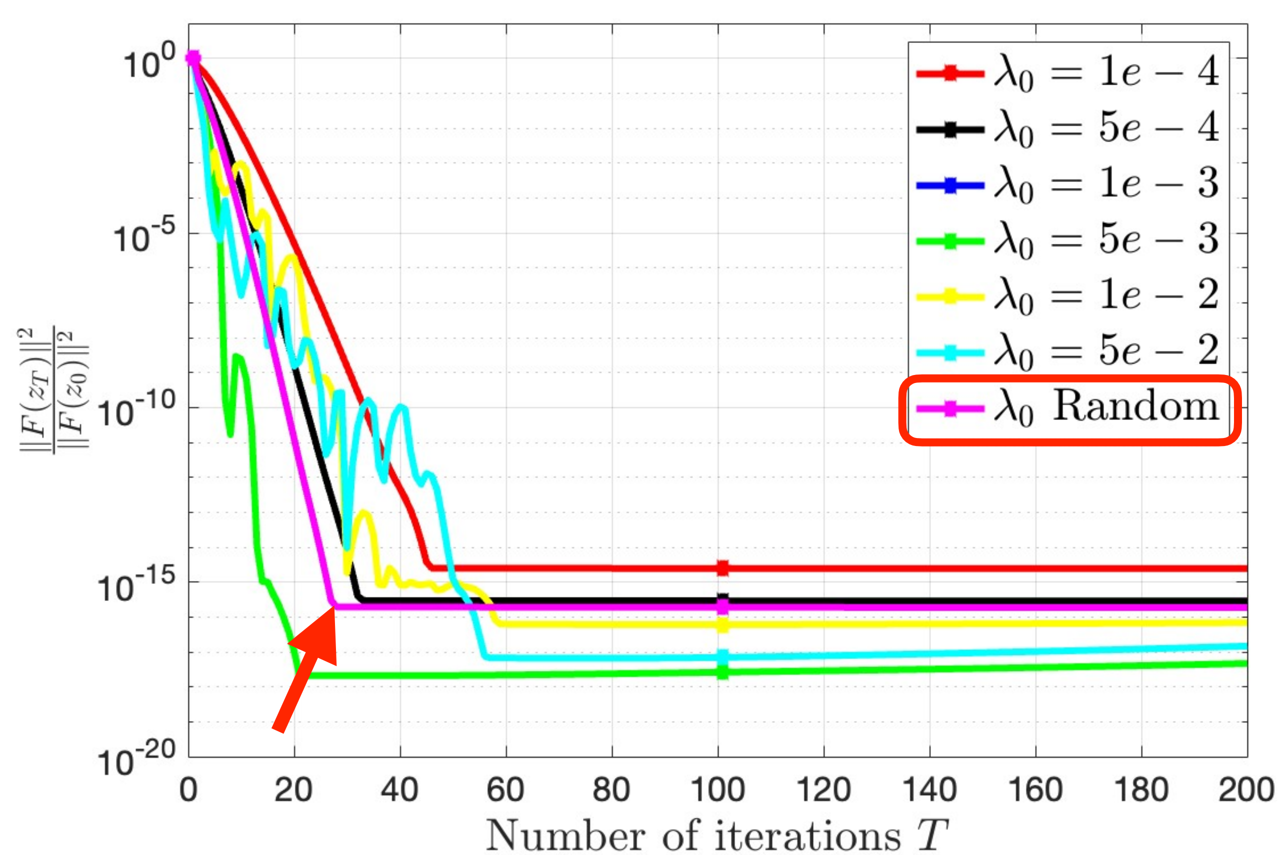


Some experiments

$$\min_{\mathbf{x}\in\mathbb{R}^n}\max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y}) = (\mathbf{A}\mathbf{x}-\mathbf{b})^\top\mathbf{y} + (L_2/6)\|\mathbf{x}\|^{\mathbf{3}}$$



Small Lipschitz constant



Large Lipschitz constant

The algorithm

Algorithm 1 Adaptive Second-order Optimistic Method

Input: Initial points $\mathbf{z}_0 = \mathbf{z}_1 \in \mathbb{R}^m \times \mathbb{R}^n$, initial parameters $\eta_0 = 0$ and $\lambda_0 > 0$
for $t = 1, \ldots, T$ **do**
  Set: $\mathbf{e}_t = \mathbf{F}(\mathbf{z}_t) - \mathbf{F}(\mathbf{z}_{t-1}) - \mathbf{F}'(\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{z}_{t-1})$
  Set the step size parameters

$$\lambda_t = \begin{cases} L_2 & \textbf{(I)} \\ \max\left\{\lambda_{t-1}, \frac{2\|\mathbf{e}_t\|}{\|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2}\right\} & \textbf{(II)} \end{cases} \qquad \eta_t = \frac{\lambda_t}{2\left(\eta_{t-1}\|\mathbf{e}_t\| + \sqrt{\eta_{t-1}^2\|\mathbf{e}_t\|^2 + \lambda_t\|\mathbf{F}(\mathbf{z}_t)\|}\right)}$$

  Update: $\mathbf{z}_{t+1} = \mathbf{z}_t - (\lambda_t \mathbf{I} + \eta_t \mathbf{F}'(\mathbf{z}_t))^{-1}(\eta_t \mathbf{F}(\mathbf{z}_t) + \eta_{t-1}\mathbf{e}_t)$
end for
return $\bar{\mathbf{z}}_{T+1} = (\sum_{t=0}^{T} \eta_t)^{-1} \sum_{t=0}^{T} \eta_t \mathbf{z}_{t+1}$

Thank you for listening!