Non-Expansive Mappings in Two-Time-Scale Stochastic Approximation

Siddharth Chandak

June 30, 2025

Department of Electrical Engineering, Stanford University

Outline

- Framework
- An Interesting Variant: Projections
- Convergence Rates
- Applications

Framework

Two-Time-Scale Iterations

- Coupled iterations updating on separate time-scales

$$x_{k+1} = x_k + \alpha_k(f(x_k, y_k) - x_k + M_{k+1})$$
$$y_{k+1} = y_k + \beta_k(g(x_k, y_k) - y_k + M'_{k+1})$$

- Want to solve $f(x, y) = x$ and $g(x, y) = y$ given noisy realizations
- M_{k+1} and M'_{k+1} are martingale difference noise sequences arising from noisy observations
- Timescales dictated by the different stepsizes α_k and β_k

Two-Time-Scale Iterations

$$\text{Faster: } x_{k+1} = x_k + \alpha_k(f(x_k, y_k) - x_k + M_{k+1})$$
$$\text{Slower: } y_{k+1} = y_k + \beta_k(g(x_k, y_k) - y_k + M'_{k+1})$$

- α_k is larger, or decays at a slower rate, e.g., $1/n^{0.6}$
- β_k is smaller, or decays at a faster rate, e.g., $1/n^{0.75}$
- Analysis
    - Faster time-scale: y_k considered quasi-static
    - Slower time-scale: x_k tracks $x^*(y_k)$, the fixed point for $f(\cdot, y_k)$

Contractive and Non-Expansive Mappings

- Contractive:
  - There exists $0 \leq \mu < 1$ such that for all x_1, x_2,

$$\|f(x_1) - f(x_2)\| \leq \mu \|x_1 - x_2\|$$

  - Equivalent to $x - f(x)$ being a strongly monotone operator[1]
- Non-expansive:
  - For all x_1, x_2,

$$\|f(x_1) - f(x_2)\| \leq \|x_1 - x_2\|$$

  - Equivalent to $x - f(x)$ being a co-coercive operator[2]

[1]contractive $\iff$ strongly monotone and Lipschitz
[2]non-expansive $\iff$ co-coercive

Our Framework

- $f(x, y)$ is μ-contractive in x
  - $\|f(x_1, y) - f(x_2, y)\| \leq \mu\|x_1 - x_2\|$, for all x_1, x_2, y
  - For each y, there exists unique $x^*(y)$ such that $f(x^*(y), y) = x^*(y)$
- $g(x^*(y), y)$ is non-expansive
  - $\|g(x^*(y_1), y_1) - g(x^*(y_2), y_2)\| \leq \|y_1 - y_2\|$
  - Assume that set of fixed points of $g(x^*(\cdot), \cdot)$ is non-empty (denote by $\mathcal{Y}^*$)

Difference from Literature

Our Framework

- $f(x, y)$ is μ-contractive in x
- $g(x^*(y), y)$ is non-expansive

Other Finite-Time Analysis Works

- $f(x, y)$ and $g(x^*(y), y)$ are contractive in x and y, respectively
- Unique fixed points and solution to $f(x^*, y^*) = x^*$ and $g(x^*, y^*) = y^*$
- Stronger convergence results

An Interesting Variant

Projected Variant

$$
\begin{aligned}
x_{k+1} &= \mathcal{P}_{\mathcal{X}}\big(x_k + \alpha_k(f(x_k, y_k) - x_k + M_{k+1})\big) \\
y_{k+1} &= y_k + \beta_k(g(x_k, y_k) - y_k + M'_{k+1})
\end{aligned}
$$

- $\mathcal{P}_{\mathcal{X}}$: Projection onto $\mathcal{X}$, a convex and bounded set
- $\mathcal{P}_{\mathcal{X}}(x) = \arg\min_{x' \in \mathcal{X}} \|x - x'\|$
- How this setup is different?
  - Relevant fixed point is $\hat{x}(y)$, the fixed point for $\mathcal{P}_{\mathcal{X}}(f(\cdot, y))$
  - $\mathcal{P}_{\mathcal{X}}(f(\cdot, y))$ still assumed to be μ-contractive
  - Now $g(\hat{x}(\cdot), \cdot)$ assumed to be non-expansive

Interesting Scenarios

- Recall that $x^*(y)$ is fixed point for $f(\cdot, y)$ and $\hat{x}(y)$ is fixed point for $\mathcal{P}_{\mathcal{X}}(f(\cdot, y))$
- There exist scenarios where $g(x^*(y), y)$ is contractive, but $g(\hat{x}(y), y)$ is non-expansive (and not contractive)
- **Implication:** Faster convergence rate analysis possible in absence of projection

Convergence Rates

Assumptions

- Functions $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are L-Lipschitz
- M_{k+1} and M'_{k+1} are martingale difference sequences with respect to $\mathcal{F}_k = \sigma(x_0, y_0, M_i, M'_i, i \leq k)$. Moreover,

$$\mathbb{E}[\|M_{k+1}\|^2 + \|M'_{k+1}\|^2 \mid \mathcal{F}_k] \leq \mathfrak{c}_2(1 + \|x_k\|^2 + \|y_k\|^2), \ \forall k \geq 0$$

- Stepsize Sequences:

$$\alpha_k = \frac{\alpha_0}{(k+1)^{\mathfrak{a}}} \text{ and } \beta_k = \frac{\beta_0}{(k+1)^{\mathfrak{b}}},$$

where $0.5 < \mathfrak{a} < \mathfrak{b} < 1$. Importantly,

$$\frac{\beta_k^2}{\alpha_k^3} \leq 1$$

Convergence Rate

Theorem

The iterates satisfy the following:

- $\mathbb{E}\left[\|x_k - x^*(y_k)\|^2\right] = \mathcal{O}\left(1/(k+1)^{\mathfrak{a}}\right)$
- $\mathbb{E}\left[\|g(x^*(y_k), y_k) - y_k\|^2\right] = \mathcal{O}\left(1/(k+1)^{1-\mathfrak{b}}\right)$
- $\|x_k - x^*(y_k)\|$ converges to zero, and the iterates y_k converge to the set $\mathcal{Y}^*$ with probability 1.

- **Remark:** The same result holds for the projected variant as well (replacing $x^*(y)$ with $\hat{x}(y)$)

Optimal Rate

- Want to maximize $1 - \mathfrak{b}$ under the constraints, $0.5 < \mathfrak{a} < \mathfrak{b}$ and $2\mathfrak{b} \geq 3\mathfrak{a}$
- Optimal Rate: $\mathcal{O}\left(1/k^{0.25-\epsilon}\right)$, achieved at $\alpha_k = 1/(k+1)^{0.5+2/3\epsilon}$ and $\beta_k = 1/(k+1)^{0.75+\epsilon}$
- Here $\epsilon > 0$ can be arbitrarily small

Comparison

Contractive-Nonexpansive	Contractive-Contractive[3]

• $\mathbb{E}\left[\|x_k - x^*(y_k)\|^2\right] = \mathcal{O}\left(\frac{1}{k^{\mathfrak{a}}}\right)$
• $\mathbb{E}\left[\|g(x^*(y_k), y_k) - y_k\|^2\right] = \mathcal{O}\left(\frac{1}{k^{1-b}}\right)$
• Optimal Rate: $\mathcal{O}\left(\frac{1}{k^{0.25-\epsilon}}\right)$
  • $\alpha_k = \mathcal{O}\left(\frac{1}{k^{0.5+2/3\epsilon}}\right)$
  • $\beta_k = \mathcal{O}\left(\frac{1}{k^{0.75+\epsilon}}\right)$
	• $\mathbb{E}\left[\|x_k - x^*(y_k)\|^2\right] = \mathcal{O}\left(\frac{1}{k^{\mathfrak{a}}}\right)$
	• $\mathbb{E}\left[\|y_k - y^*\|^2\right] = \mathcal{O}\left(\frac{1}{k^{\mathfrak{a}}}\right)$
	• Optimal Rate: $\mathcal{O}\left(\frac{1}{k}\right)$
	  • $\alpha_k = \mathcal{O}\left(\frac{1}{k}\right)$
	  • $\beta_k = \mathcal{O}\left(\frac{1}{k}\right)$

[3] Chandak, Siddharth. "$O(1/k)$ Finite-Time Bound for Non-Linear Two-Time-Scale Stochastic Approximation." *arXiv:2504.19375 (2025)*.

Applications

Linear Stochastic Approximation

- Want to solve $A_{11}x + A_{12}y = b_1$ and $A_{21}x + A_{22}y = b_2$.
- Receive unbiased estimates $\tilde{A}_{ij}^{(k+1)}$ and $\tilde{b}_i^{(k+1)}$

$$x_{k+1} = x_k + \alpha_k \left(\tilde{b}_1^{(k+1)} - \tilde{A}_{11}^{(k+1)} x_k - \tilde{A}_{12}^{(k+1)} y_k\right)$$
$$y_{k+1} = y_k + \beta_k \left(\tilde{b}_2^{(k+1)} - \tilde{A}_{21}^{(k+1)} x_k - \tilde{A}_{22}^{(k+1)} y_k\right).$$

- Define $\Delta = A_{22} - A_{21}A_{11}^{-1}A_{12}$

Corollary

If $-A_{11}$ is negative definite and $-\Delta$ is (non-zero) negative semidefinite, then

$$\mathbb{E}\left[\|A_{11}x + A_{12}y - b_1\|^2 + \|A_{21}x + A_{22}y - b_2\|^2\right] = \mathcal{O}(1/k^{0.25-\epsilon})$$

Strongly Concave-Convex Minimax Optimization

- Minimax Optimization Problem: $\min_{y \in \mathbb{R}^{d_2}} \max_{x \in \mathbb{R}^{d_1}} H(x, y)$
- Two-Time-Scale Stochastic Gradient Descent Ascent (TTSGDA):

$$x_{k+1} = x_k + \alpha_k(\nabla_x H(x_k, y_k) + M_{k+1})$$
$$y_{k+1} = y_k + \beta_k(-\nabla_y H(x_k, y_k) + M'_{k+1}).$$

- Define $\Phi(y) = \max_{x \in \mathbb{R}^{d_1}} H(x, y)$

Corollary

If $H(x, y)$ is smooth, strongly concave in x and convex in y, then the iterates converge to the set of saddle points of $H(x, y)$ a.s., and $\mathbb{E}[\|\nabla \Phi(y)\|^2] = \mathcal{O}(1/k^{0.25-\epsilon})$.

Constrained Optimization

$$\text{maximize } G(x)$$
$$\text{subject to:}$$
$$H_i(x) \leq 0, i = 1, \ldots, m$$
$$Ax = b_0.$$

- $\mathcal{X} = \{x \mid H_i(x) \leq 0, i = 1, \ldots, m\}$
- $Ax = b_0$ represents the additional linear constraint
- Two-time-scale Lagrangian optimization:

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\big(x_k + \alpha_k(\nabla G(x_k) - A^T\lambda_k + M_{k+1})\big)$$
$$\lambda_{k+1} = \lambda_k + \beta_k(Ax_k - b_0).$$

- Applications:
  - Distributed optimization with separate local and global constraints
  - Game control or generalized Nash equilibrium problems (GNEP)

Constrained Optimization

- Scenario where projection leads to non-expansiveness
- Consider no inequality constraints

$$\begin{aligned} &\text{maximize } \; G(x) \\ &\text{subject to:} \\ &\qquad \cancel{H_i(x) \leq 0, i = 1, \ldots, m} \\ &\qquad Ax = b_0. \end{aligned}$$

$$\begin{aligned} x_{k+1} &= \cancel{\mathcal{P}_{\mathcal{X}}}\big(x_k + \alpha_k(\nabla G(x_k) - A^T \lambda_k + M_{k+1})\big) \\ \lambda_{k+1} &= \lambda_k + \beta_k(Ax_k - b_0). \end{aligned}$$

- This is an example of contractive-contractive iterations

Constrained Optimization

- Coming back to our original constrained problem with the following two-time-scale iterations:

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\big(x_k + \alpha_k(\nabla G(x_k) - A^T\lambda_k + M_{k+1})\big)$$
$$\lambda_{k+1} = \lambda_k + \beta_k(Ax_k - b_0).$$

Theorem

For continuously differentiable and convex $H_i(\cdot)$ and if there exists $x \in \mathrm{int}(\mathcal{X})$ such that $Ax = b_0$, then

$$\mathbb{E}[\|A\hat{x}(\lambda_k) - b_0\|^2] = \mathcal{O}(1/k^{0.25-\epsilon})$$

Thank You!

Thank You!

The talk was primarily based on

- Chandak, Siddharth, "Non-Expansive Mappings in Two-Time-Scale Stochastic Approximation: Finite-Time Analysis." *arXiv:2501.10806* (2025).

with some results from

- Chandak, Siddharth, "$O(1/k)$ Finite-Time Bound for Non-Linear Two-Time-Scale Stochastic Approximation." *arXiv:2504.19375* (2025).
- Chandak, Siddharth, Ilai Bistritz, and Nicholas Bambos, "Learning to Control Unknown Strongly Monotone Games." *arXiv:2407.00575* (2024).