Non-Nesterov Acceleration Methods in First-Order Optimization

Ernest K. Ryu

Department of Mathematical Sciences
Seoul National University

SIAM Conference on Optimization
May 31, 2023

Acceleration of first-order convex minimization

Consider

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)$$

where f is L-smooth convex.

Gradient descent

$$x_{k+1} = x_k^+ \stackrel{\text{def}}{=} x_k - \frac{1}{L} f(x_k)$$

converges with the rate $f(x_k) - f_\star \leq \mathcal{O}(1/k)$.

Nesterov's celebrated fast gradient method (FGM):

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}}(x_k^+ - x_{k-1}^+),$$

with $\theta_0 = 1$, $\theta_i = \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2}$ for $i = 1\ldots, N$, converges with the accelerated rate $f(x_k) - f_\star \leq \mathcal{O}(1/k^2)$.

Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, *Proceedings of the USSR Academy of Sciences*, 1983.

OGM beats FGM

Surprisingly, it was discovered (via the PEP) that FGM is suboptimal.

The optimized gradient method (OGM)[#]:

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}}(x_k^+ - x_{k-1}^+) + \frac{\theta_k}{\theta_{k+1}}(x_k^+ - x_k)$$

with $\theta_i = \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2}$ for $i = 1 \ldots, N-1$, and $\theta_N = \frac{1+\sqrt{1+8\theta_{N-1}^2}}{2}$ beats FGM by a factor of 2.

Recently, several new acceleration mechanisms, distinct from Nesterov's, were discovered using the computer-assisted methodology called the PEP.

[#]Drori and Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach. *MPA*, 2014.

[#]Kim and Fessler, Optimized first-order methods for smooth convex minimization, *MPA*, 2016.

Optimal methods via PEP

Conceptually, the problem of finding the best first-order method is an optimization problem.

Surprisingly, this optimization problem can be posed as a finite-dimensional convex SDP[#] or a non-convex QCQP[†].

The following new acceleration mechanisms were designed with the PEP.

The BnB-PEP[†] is the most recent development of the tool, and this vastly expands the applicability of the computer-assisted methodology. (✓Talk here at SIOpt.)

[#] Drori and Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach. *MPA*, 2014.

[#] Taylor, Hendrickx, and Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods, *MPA*, 2017.

[†] Das Gupta, Van Parys, and **Ryu**, Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods, *MPA*, 2023.

Outline

Smooth convex-concave minimax optimization

Consider

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \underset{y\in\mathbb{R}^m}{\text{maximize}}\quad \mathbf{L}(x,y),$$

where $\mathbf{L}$ is convex-concave and R-smooth. Recently, minimax optimization has gained popularity in machine learning.

$(x_\star, y_\star)$ solves the minimax problem if it is a saddle point, i.e., if

$$\mathbf{L}(x_\star, y) \le \mathbf{L}(x_\star, y_\star) \le \mathbf{L}(x, y_\star), \qquad \forall\, x \in \mathbb{R}^n,\, y \in \mathbb{R}^m.$$

Saddle operator is

$$\mathbf{G}(x,y) \triangleq \begin{bmatrix} \nabla_x \mathbf{L}(x,y) \\ -\nabla_y \mathbf{L}(x,y) \end{bmatrix}.$$

$\mathbf{L}$ is R-smooth of $\mathbf{G}$ is R-Lipschitz continuous. $z = (x,y)$ is a saddle point of $\mathbf{L}$ if and only if $\mathbf{G}(z) = 0$.

Question) Can we accelerate first-order minimax algorithms?

Classical results in minimax optimization

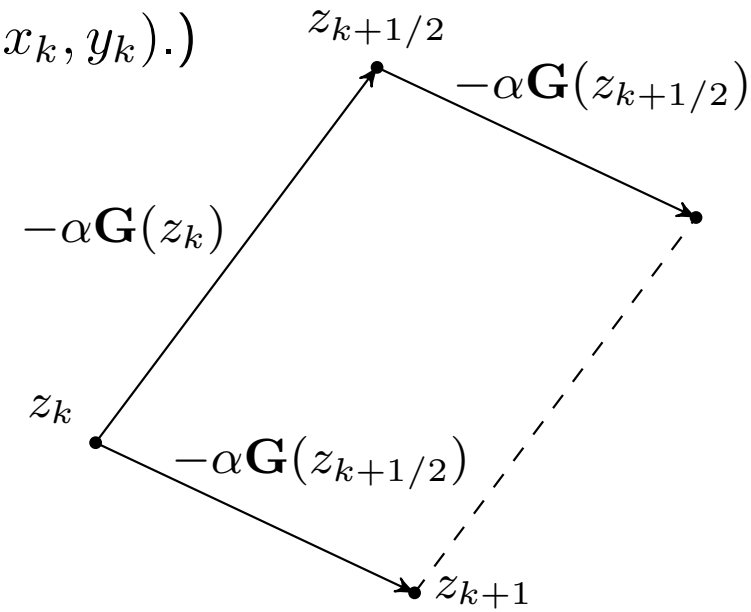
Analogue of gradient descent (simultaneous gradient descent-ascent)

$$z_{k+1} = z_k - \alpha\, \mathbf{G}(z_k),$$

does not converge in general. (Write $z_k = (x_k, y_k)$.)

Extragradient (EG) algorithm[1]

$$z_{k+1/2} = z_k - \alpha\, \mathbf{G}(z_k)$$
$$z_{k+1} = z_k - \alpha\, \mathbf{G}(z_{k+1/2})$$



does converge.

Theorem (Informal)

EG and several other known methods exhibit the rate

$$\min_{i=0,\ldots,k} \|\nabla \mathbf{L}(z_i)\|^2 \leq \mathcal{O}\left(\frac{R^2 \|z_0 - z_\star\|^2}{k}\right).$$

[1] Korpelevich, The extragradient method for finding saddle ..., *Matekon*, 1976.

Extra anchored gradient (EAG) algorithm

$$z_{k+1/2} = z_k + \frac{1}{k+2}(z_0 - z_k) - \alpha\,\mathbf{G}(z_k)$$
$$z_{k+1} = z_k + \frac{1}{k+2}(z_0 - z_k) - \alpha\,\mathbf{G}(z_{k+1/2})$$

$\alpha > 0$ is the *step-size* and $\frac{1}{k+2}$ are *anchoring coefficients*.
Anchor term pulls z_k towards the initial point z_0.

Theorem

With $\alpha \leq \frac{1}{8R}$, EAG exhibits the rate

$$\|\nabla \mathbf{L}(z_k)\|^2 \leq \mathcal{O}\left(\frac{R^2\|z_0 - z_\star\|^2}{k^2}\right).$$

Yoon and **Ryu**, Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm, *ICML long talk*, 2021.

EAG is optimal up to a constant

Theorem (Informal)

EAG is optimal up to a constant among algorithms satisfying:

$$x_i \in x_0 + \operatorname{span}\{\nabla_x \mathbf{L}(x_0, y_0), \ldots, \nabla_x \mathbf{L}(x_{i-1}, y_{i-1})\}$$
$$y_i \in y_0 + \operatorname{span}\{\nabla_y \mathbf{L}(x_0, y_0), \ldots, \nabla_y \mathbf{L}(x_{i-1}, y_{i-1})\}$$

for $i = 1, \ldots, k$.

Yoon and **Ryu**, Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm, *ICML long talk*, 2021.

Related follow-up work

- Lee, D. Kim, Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems, *NeurIPS*, 2021. ✓
- Tran-Dinh, Luo, Halpern-type accelerated and splitting algorithms for monotone inclusions, *arXiv*, 2021.
- Yoon, **Ryu**, Accelerated minimax algorithms flock together, *arXiv*, 2022. ✓
- Bot, Csetnek, Nguyen, Fast OGDA in continuous and discrete time, *arXiv*, 2022. ✓
- SedImayer, Nguyen, Bot, A fast optimistic method for monotone variational inequalities, *ICML*, 2023.

✓: Talks given here in SIOpt.

Outline

Acceleration for minimax optimization

Acceleration for fixed-point iterations

Acceleration for composite optimization

Acceleration for making gradients small

Fixed-point iteration

Fixed-point iteration with $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ computes

$$x_{k+1} = \mathbb{T}x_k$$

with some starting point $x_0 \in \mathbb{R}^n$.

Surprisingly, the classical fixed-point iteration is suboptimal.

Question) What is the optimal (accelerated) iteration complexity of fixed-point iterations?

Accelerated fixed-point iteration

Optimal Contractive Halpern (OC-Halpern) was discovered by the PEP:

$$y_k = \left(1 - \frac{1}{\varphi_k}\right) \mathbb{T} y_{k-1} + \frac{1}{\varphi_k} y_0 \qquad \text{(OC-Halpern)}$$

where $\mathbb{T}$ is $1/\gamma$-contractive, $\varphi_k = \sum_{i=0}^{k} \gamma^{2i}$, and y_0 is a starting point.

Theorem

(OC-Halpern) *exhibits the rate*

$$\|y_N - \mathbb{T} y_N\|^2 \leq \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^{N} \gamma^k}\right)^2 \|y_0 - y_\star\|^2.$$

Faster than plain fixed-point iteration. When $\gamma = 1$, the rate

$$\|y_N - \mathbb{T} y_N\|^2 \leq \mathcal{O}(1/N^2)$$

is faster than the $\mathcal{O}(1/N)$ rate for plain (KM) fixed-point iteration.

$\mathcal{O}(1/N^2)$ rate due to Lieder and Kim. Rate for $\gamma > 1$ due to Park and Ryu.

Lieder, On the convergence rate of the Halpern-iteration. *OPTL*, 2021.
Kim, Accelerated proximal point method ..., *MPA*, 2021.
Park and **Ryu**, Exact optimal accelerated complexity for ..., *ICML long talk*, 2022.

Optimality

Theorem (Informal)

OC-Halpern is exactly among algorithms satisfying:

$$y_k \in y_0 + \operatorname{span}\{y_0 - \mathbb{T}y_0, y_1 - \mathbb{T}y_1, \ldots, y_{k-1} - \mathbb{T}y_{k-1}\}$$

for $k = 1, \ldots, N$.

Diakonikolas, Halpern iteration for near-optimal and parameter-free ..., *COLT*, 2020.
Park and **Ryu**, Exact optimal accelerated complexity for ..., *ICML long talk*, 2022.

Anchored value iteration: Upper bound

Let T be the Bellman optimality operator. *Anchored Value Iteration* is

$$U^k = \frac{1}{\varphi_k} U^0 + \left(1 - \frac{1}{\varphi_k}\right) T U^{k-1}, \qquad \text{(Anc-VI)}$$

same as OC-Halpern. (T is γ-contractive in the $\|\cdot\|_\infty$-norm.)

Theorem

Let $0 < \gamma < 1$ be the discount factor. Let T be the Bellman optimality operator. If $U^0 \le TU^0$, then Anc-VI exhibits the rate

$$\begin{aligned} \left\|TU^k - U^k\right\|_\infty &\le \frac{\left(\gamma^{-1} - \gamma\right)\left(1 + \gamma - \gamma^{k+1}\right)}{\left(\gamma^{k+1}\right)^{-1} - \gamma^{k+1}} \left\|U^0 - U^\star\right\|_\infty \\ &= \left(\frac{1}{k+1} + \frac{k}{k+1}\epsilon + O(\epsilon^2)\right) \left\|U^0 - U^\star\right\|_\infty \end{aligned}$$

where $\gamma = 1 - \epsilon$.

Lee and **Ryu**, Accelerating value Iteration with anchoring, *arXiv*, 2023.

Anchored value iteration: Lower bound

Theorem

Let $k \geq 0$, $n \geq k+2$, $0 < \gamma \leq 1$, and $U^0 \in \mathbb{R}^n$. Then there exists an MDP with $|\mathcal{S}| = n$ and $|\mathcal{A}| = 1$ such that T has a fixed point $U^\star$ satisfying $U^0 \leq U^\star$ and

$$\left\|TU^k - U^k\right\|_\infty \geq \frac{\gamma^k}{\sum_{i=0}^k \gamma^i} \left\|U^0 - U^\star\right\|_\infty$$

for any iterates $\{U^i\}_{i=0}^k$ satisfying the span condition

$$U^i \in U^0 + \operatorname{span}\{TU^0 - U^0, TU^1 - U^1, \ldots, TU^{i-1} - U^{i-1}\}.$$

Since

$$\frac{\gamma^k}{\sum_{i=0}^k \gamma^i} \leq \frac{\left(\gamma^{-1} - \gamma\right)\left(1 + \gamma - \gamma^{k+1}\right)}{\left(\gamma^{k+1}\right)^{-1} - \gamma^{k+1}} \leq \frac{4\gamma^k}{\sum_{i=0}^k \gamma^i} \qquad \forall\, 0 < \gamma < 1,$$

upper bound is optimal up to a constant of factor 4.

Lee and **Ryu**, Accelerating value Iteration with anchoring, *arXiv*, 2023.

Outline

Acceleration for minimax optimization

Acceleration for fixed-point iterations

Acceleration for composite optimization

Acceleration for making gradients small

Composite optimization and FISTA

Composite minimization

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x)+h(x),$$

f is L-smooth convex and h is convex proximable.

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA):

$$y_{i+1} = \mathbf{Prox}_{\frac{1}{L}h}\left(x_i - \frac{1}{L}\nabla f(x_i)\right),$$
$$x_{i+1} = y_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}\left(y_{i+1} - y_i\right),$$

with $\theta_0 = 1$, $\theta_i = \frac{\left(1+\sqrt{1+4\theta_{i-1}^2}\right)}{2}$ for $i = 1,\ldots,N-1$. Convergence rate:

$$f(y_N) + h(y_N) - f(x_\star) - h(x_\star) \le \frac{L\|x_0 - x_\star\|^2}{2\theta_{N-1}^2} \le \frac{2L\|x_0 - x_\star\|^2}{(N+1)^2}.$$

Beck, Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SISC*, 2009

Exact optimal method: OptISTA

Optimal Iterative Shrinkage Thresholding Algorithm

$$
\begin{aligned}
y_{i+1} &= \mathbf{Prox}_{\frac{\gamma_i}{L}h}\left(y_i - \frac{\gamma_i}{L}\nabla f(x_i)\right), \\
z_{i+1} &= x_i + \frac{1}{\gamma_i}\left(y_{i+1} - y_i\right), \qquad \text{(OptISTA)} \\
x_{i+1} &= z_{i+1} + \frac{\theta_i - 1}{\theta_{i+1}}\left(z_{i+1} - z_i\right) + \frac{\theta_i}{\theta_{i+1}}\left(z_{i+1} - x_i\right),
\end{aligned}
$$

with $\gamma_i = \frac{2\theta_i}{\theta_N^2}\left(\theta_N^2 - 2\theta_i^2 + \theta_i\right)$,
$\theta_i = \frac{1+\sqrt{1+4\theta_{i-1}^2}}{2}$ for $i = 1\ldots, N-1$, and $\theta_N = \frac{1+\sqrt{1+8\theta_{N-1}^2}}{2}$.

Jang, Das Gupta, **Ryu**, Computer-assisted design of accelerated composite optimization methods: OptISTA, *arXiv*, 2023.

Exact optimal method: OptISTA

OptISTA improves upon the rate of FISTA by a factor of 2:

Theorem

OptISTA exhibits the rate

$$f(y_N) + h(y_N) - f(x_\star) - h(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)} \leq \frac{L\|x_0 - x_\star\|^2}{(N+1)^2},$$

OptISTA is exactly optimal:

Theorem

Let $L > 0$, $R > 0$, $N > 0$, and $d \geq N + 1$. Under an appropriate span condition, there is an f and h such that

$$f(x_N) + h(x_N) - f(x_\star) - h(x_\star) \geq \frac{L\|x_0 - x_\star\|^2}{2(\theta_N^2 - 1)}.$$

Jang, Das Gupta, **Ryu**, Computer-assisted design of accelerated composite optimization methods: OptISTA, *arXiv*, 2023.

Outline

Making gradients small fast for convex functions

Consider

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x)$$

where f is L-smooth convex. Let us measure suboptimality by $\|\nabla f(x)\|^2$, rather than $f(x) - f_\star$.

Gradient descent exhibits the rate[2]

$$\|\nabla f(x_k)\|^2 \leq \mathcal{O}(1/k^2)$$

FGM exhibits the rate[3]

$$\min_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 \leq \mathcal{O}(1/k^3)$$

Nemirovsky establishes the lower bound[4]

$$\min_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 \geq \Theta(1/k^4).$$

Question) What is the optimal accelerated rate for making gradients of convex functions small?

---

[2] Taylor and Bach, Stochastic first-order methods: non-asymptotic ..., *COLT*, 2019.
[3] Shi, Du, Su, and Jordan, Acceleration via symplectic ..., *MPA*, 2019.
[4] Nemirovsky, Information-based complexity of linear ..., *J. Complexity*, 1992.

OGM-G: $\mathcal{O}((f(x_0) - f_\star)/K^2)$ rate

OGM-G:

$$x_{k+1} = x_k^+ + \frac{(\theta_k - 1)(2\theta_{k+1} - 1)}{\theta_k(2\theta_k - 1)}(x_k^+ - x_{k-1}^+) + \frac{2\theta_{k+1} - 1}{2\theta_k - 1}(x_k^+ - x_k)$$

where $x^+ = x - \frac{1}{L}\nabla f(x)$, $x_{-1}^+ = x_0$, $\theta_K = 1$, and $\theta_k^2 - \theta_k = \theta_{k+1}^2$.

OGM-G, discovered with the PEP, exhibits the rate

$$\|\nabla f(x_K)\|^2 \leq \mathcal{O}\left(\frac{f(x_0) - f_\star}{K^2}\right).$$

Kim and Fessler, Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions, *JOTA*, 2021.

FGM+OGM-G: $\mathcal{O}(\|x_0 - x_\star\|^2/K^4)$ rate

FGM+OGM-G: From x_0 run K iterations of FGM. Continue with OGM-G and run K iterations. Concatenated method exhibits the rate

$$\|\nabla f(x_{2K})\|^2 \leq \mathcal{O}(\|x_0 - x_\star\|^2/K^4)$$

FGM: $\mathcal{O}(1/K^2)$ rate on $\big(\|x_0 - x_\star\|^2 \mapsto f(x_K) - f(x_\star)\big)$.

OGM-G: $\mathcal{O}(1/K^2)$ rate on $\big(f(x_0) - f(x_\star) \mapsto \|\nabla f(x_K)\|^2\big)$.

FGM+OGM-G: $\mathcal{O}(1/K^4)$ rate on $\big(\|x_0 - x_\star\|^2 \mapsto \|\nabla f(x_{2K})\|^2\big)$.

Nesterov, Gasnikov, Guminov, and Dvurechensky, Primal–dual accelerated gradient methods ..., *Optimization Methods and Software*, 2020.

Prox-grad setup

Consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad F(x) := f(x) + g(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex L-smooth and g is proximable.

Prox-grad step notation:

$$x^{\oplus} = \underset{y\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ f(x) + \langle \nabla f(x), y - x\rangle + g(y) + \frac{L}{2}\|y - x\|^2 \right\}.$$

FISTA-G

A novel method, FISTA-G:

$$x_{k+1} = x_k^{\oplus} + \frac{\varphi_{k+1} - \varphi_{k+2}}{\varphi_k - \varphi_{k+1}}(x_k^{\oplus} - x_{k-1}^{\oplus}),$$

where $x_{-1}^{\oplus} := x_0$, $\varphi_{K+1} = 0$, $\varphi_K = 1$, and

$$\varphi_k = \frac{\varphi_{k+2}^2 - \varphi_{k+1}\varphi_{k+2} + 2\varphi_{k+1}^2 + (\varphi_{k+1} - \varphi_{k+2})\sqrt{\varphi_{k+2}^2 + 3\varphi_{k+1}^2}}{\varphi_{k+1} + \varphi_{k+2}}.$$

Theorem

FISTA-G exhibits the rate

$$\min \|\partial F(x_K^{\oplus})\|^2 \leq \frac{264L(F(x_0) - F_\star)}{(K+2)^2}.$$

Lee, Park, and **Ryu**, A geometric structure of acceleration and its role in making gradients small fast, *NeurIPS*, 2021.

FISTA+FISTA-G: $\mathcal{O}((\|x_0 - x_\star\|^2)/K^4)$ rate

Corollary

FISTA+FISTA-G's final iterate x_{2K} exhibits the rate

$$\min \|\partial F(x_{2K}^{\oplus})\|^2 \leq \mathcal{O}\left(\frac{L^2\|x_0 - x_\star\|^2}{K^4}\right).$$

Lee, Park, and **Ryu**, A geometric structure of acceleration and its role in making gradients small fast, *NeurIPS*, 2021.

Super-FISTA-G

Constant factor 264 is improved to 50 by *Super-FISTA-G*

$$x_{k+1} = x_k^{\oplus,4} + \frac{(K-k+1)(2K-2k-1)}{(K-k+3)(2K-2k+1)}\left(x_k^{\oplus,4} - x_{k-1}^{\oplus,4}\right) + \frac{(4K-4k-1)(2K-2k-1)}{6(K-k+3)(2K-2k+1)}\left(x_k^{\oplus,4} - x_k\right)$$

for $k = 0, \ldots, K-2$

$$x_K = x_{K-1}^{\oplus,4} + \frac{3}{10}\left(x_{K-1}^{\oplus,4} - x_{K-2}^{\oplus,4}\right) + \frac{3}{40}\left(x_{K-1}^{\oplus,4} - x_{K-1}\right) \qquad \text{(SFG)}$$

where we use the α-proximal gradient step

$$y^{\oplus,\alpha} = \operatorname*{argmin}_{z\in\mathbb{R}^n}\left(f(y) + \langle \nabla f(y), z-y\rangle + g(z) + \frac{\alpha L}{2}\|z-y\|^2\right) = \operatorname{Prox}_{\frac{g}{\alpha L}}\left(y - \frac{1}{\alpha L}\nabla f(y)\right)$$

Theorem

SFG exhibits the rate

$$\min_{v\in\partial F(x_K^{\oplus,4})} \|v\|^2 \le \frac{50L}{(K+2)(K+3)}\left(F(x_0) - F_\star\right).$$

Kim, Ozdaglar, Park, and **Ryu**, Time-reversed dissipation induces duality between minimizing gradient norm and function value, *arXiv*, 2023.

Conclusion

With the aid of the PEP, the field has discovered several new acceleration mechanisms, exciting developments in convex optimization theory.

Future outlook #1: There are many interesting problems in optimization theory that the BnB-PEP will finally empower us to tackle.

Future outlook #2: Using the PEP to analyze algorithms outside of optimization, e.g., numerical analysis and reinforcement learning.