A regret minimization approach to fixed point iterations

Joon Kwon

INRAE

January 8th, 2025

Mokameeting
INRIA Paris

Introduction

Core result

A link between:

- Regret minimization
  a sequential decision framework with known links with first-order optimization.
- Fixed point iterations
  iterative methods for solving fixed point problems.

Application

- Define novel fixed point iterations
  based on regret minimizing methods, with guarantees transposed from regret bounds.
- In particular, AdaGrad-based iterations
  with adaptive guarantees.

Summary

- Reminder: Fixed point problems and iterations
  a very general approach for designing iterative methods
- Reminder: Regret minimization
  a classical sequential decision problem
- Link between regret minimization and fixed point iterations
  a core lemma
- Reminder: AdaGrad for regret minimization and optimization
  with an adaptive character and good properties, both theoretical and practical
- AdaGrad-based fixed point iterations
  novel iterations with adaptive guarantees

Fixed point problems

Let $F : X \to X$ where $X \subset V$ (vector space).

Goal

Find $x_* \in X$ such that $F(x_*) = x_*$.

Numerous applications

- Linear systems (Richardson, Gauss-Seidel, Jacobi)
- Ordinary/partial differential equations
- Dynamic programming and reinforcement learning (Q-learning)
- Optimization (Sinkhorn, gradient descent, forward-backward, ADMM, Chambolle-Pock, etc.)
- Statistics (EM algorithm)

Examples

EM Algorithm (Dempster et al., 1977)

for latent variable models

$(Y, Z) \sim p_\theta(y, z)$, Y observed, Z latent.

$$\theta_{t+1} = \arg\max_{\theta \in \Theta} \mathbb{E}_{Z \sim p_{\theta_t}(\,\cdot\,|Y)} \left[\log p_\theta(Y, Z)\right],$$

looks for a fixed point of operator:

$$\theta \mapsto \arg\max_{\theta' \in \Theta} \mathbb{E}_{Z \sim p_\theta(\,\cdot\,|Y)} \left[\log p_{\theta'}(Y, Z)\right].$$

Sinkhorn's algorithm (Cuturi, 2013)

for entropic optimal transport

$\varepsilon > 0$, $a \in \Delta_m$, $b \in \Delta_n$, $U(a, b)$ transport plans, C cost matrix, $K = e^{-C/\varepsilon}$, H negative entropy.

$$\min_{P \in U(a,b)} \left\{\langle P, C\rangle - \varepsilon H(P)\right\}.$$

Equivalent to finding $u \in \mathbb{R}_+^m$ and $v \in \mathbb{R}_+^n$ such that:

$$u = \frac{a}{Kv} \quad \text{and} \quad v = \frac{b}{K^\top u}.$$

Corresponding fixed point iterations:

$$u_{t+1} = \frac{a}{Kv_t} \quad \text{and} \quad v_{t+1} = \frac{b}{K^\top u_{t+1}}.$$

Fixed point iterations with contractive operators

Theorem (Banach, 1922)

Let (X, d) be a complete metric space, $F : X \to X$ a L-Lipschitz map with $0 \leqslant L < 1$. Then,

- *F admits a unique fixed point $x_* \in X$,*
- *for all $x_1 \in X$ and*

$$x_{t+1} = F(x_t), \qquad t \geqslant 1,$$

it holds that

$$d(x_T, x_*) \leqslant L^{T-1} d(x_1, x_*), \qquad T \geqslant 1.$$

(geometric convergence)

Example: Linear systems

Let $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$.

(Richardson iteration, 1910)

$$Ax = b$$

$$\Updownarrow$$

$$\underbrace{x + (b - Ax)}_{=: F(x)} = x$$

$$\Updownarrow$$

$$\underbrace{x + \gamma(b - Ax)}_{=: F_\gamma(x)} = x$$

$$\begin{aligned} x_{t+1} &= F(x_t) \\ &= x_t + (b - Ax_t) \end{aligned}$$

$$\begin{aligned} x_{t+1} &= F_\gamma(x_t) \\ &= x_t + \gamma(b - Ax_t) \end{aligned}$$

- Needs $\gamma \neq 0$ such that $F_\gamma = I + \gamma(F - I)$ is a contraction
- There are other types of iterations for specific classes of matrices (Jacobi, Gauss–Seidel, etc.).

Fixed point iterations with nonexpansive operators

Let $F : X \to X$ be nonexpansive (i.e. 1-Lipschitz).

- F may have no fixed point.
  for e.g. a translation
- Even if a fixed point exists, iteration $x_{t+1} = F(x_t)$ may not converge.
  for e.g. a rotation

Krasnoselskii–Mann iterations (1953)

Assume that a fixed point x_* exists, that X is convex. Let $x_1 \in X$ and

$$x_{t+1} = \frac{x_t + F(x_t)}{2}, \qquad t \geqslant 1.$$

Theorem (Baillon–Bruck, 1996; Cominetti–Sotto–Vaisman, 2014)
In finite dimension, $(x_t)_{t \geqslant 1}$ converges to a fixed point and

$$\left\| F(x_T) - x_T \right\|_2 \leqslant \frac{\left\| x_1 - x_* \right\|_2}{2\sqrt{\pi T}}, \qquad T \geqslant 1.$$

What if F is *not* nonexpansive?

- For $\gamma \neq 0$,
$$F \quad \text{and} \quad F_\gamma := I + \gamma(F - I)$$
have the same fixed points.
- If F_γ is nonexpansive for some $\gamma \neq 0$, KM with F_γ guarantees
$$\|F(x_T) - x_T\|_2 \leqslant \frac{\|x_1 - x_*\|_2}{2\gamma\sqrt{\pi T}}.$$
- Ideally, we want the largest such γ.
- But even if a such γ exists, it may be unknown.
- Related to the choice of step-size (aka learning rate) in optimization and ML/DL.

Example: First-order optimization

$$\min_{x \in \mathbb{R}^d} f(x) \qquad (f \text{ differentiable})$$

$$\nabla f(x) = 0$$
$$\Updownarrow$$
$$x - \gamma \nabla f(x) = x$$
$$(\gamma \neq 0)$$

Gradient descent

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

- Needs $\gamma \neq 0$ such that $I - \gamma \nabla f$ is contractive or nonexpansive.

Theorem (Baillon–Haddad, 1977)
If f is convex and ∇f is L-Lipschitz,

$$I - \gamma \nabla F \quad \textit{is nonexpansive for all } 0 < \gamma < 2/L.$$

- In practice, L may be unknown and difficult to estimate.
  for e.g. logistic regression
- Gradient descent is very sensitive to γ
  Small γ gives slow convergence, does not converge for large γ.
- Well-known to MD/DL practionniers
  Tuning is compuationnaly heavy

Regret minimization

Sequential decision problem involving a Player against Nature

Introduced by (Hannan, 1957)

Online linear optimization (OLO)

(Zinkevich, 2003)

$$\mathcal{X} \subset \mathbb{R}^d \quad \text{convex compact,} \qquad \mathcal{U} \subset \mathbb{R}^d$$

For $t \geqslant 1$,

- Player chooses $x_t \in \mathcal{X}$
- Nature chooses $u_t \in \mathcal{U}$
- Player gets payoff $\langle u_t, x_t \rangle$

$$\text{Regret} = \frac{1}{T}\left(\max_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle u_t, x \rangle - \sum_{t=1}^{T} \langle u_t, x_t \rangle \right)$$

$$= \max_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^{T} \langle u_t, x - x_t \rangle$$

- If $\mathcal{U}$ is bounded, possible to minimize the regret as $O(1/\sqrt{T})$.

Example of regret minimizing algorithms

Online gradient descent

(Zinkevich, 2003)

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t + \gamma_t u_t), \qquad t \geqslant 1.$$

Online mirror descent

(Shalev-Shwartz, 2007)

with squared Mahalanobis distances

$$x_{t+1} = \Pi_{\mathcal{X},B}(x_t + \gamma_t B^{-1} u_t), \qquad t \geqslant 1.$$

Exponential weights algorithm

(Littlestone–Warmuth, 1994)

$$\mathcal{X} = \Delta_d = \left\{ x \in \mathbb{R}_+^d,\ \sum_{i=1}^d x_i = 1 \right\}$$

$$x_t = \left(\frac{\exp\left(\eta_t \sum_{s=0}^{t-1} u_{s,i}\right)}{\sum_{j=1}^d \exp\left(\eta_t \sum_{s=0}^{t-1} u_{s,j}\right)} \right)_{1 \leqslant i \leqslant d}, \quad t \geqslant 0.$$

Links between regret minimization and other problems

Regret can be used as a theoretical tool to define and analyze algorithms is various problems.

- First-order optimization

  Gradient descent, mirror descent (Nemirovsky–Yudin, 1983), dual averaging (Nesterov, 2009), Nesterov's acceleration (1983), etc.

- Two-player zero-sum games

  Regret matching (Hart–Mas-Colell, 2000), counterfactual regret minimization (Zinkevich, 2007), first superhuman poker algorithm (Tammelin et al., 2015).

$$\max_{x \in \Delta_m} \min_{y \in \Delta_n} \langle x, Ay \rangle \qquad (A \in \mathbb{R}^{m \times n})$$

- Variational inequalities with Lipschitz monotone operators

  Extragradient (Korpelevich, 1976), mirror-prox (Nemirovsky, 2004), dual extrapolation (Nesterov, 2007).

$$\max_{x \in X} \langle G(x_*), x - x_* \rangle \geqslant 0.$$

A link between regret minimization and fixed point problems

From now on, $X \subset \mathbb{R}^d$ is nonempty and convex, $F : X \to X$, and $x_* \in X$ a fixed point of F.

Lemma (K., 2025)

Let $\gamma > 0$ and assume that $F_\gamma = I + \gamma(F - I)$ is nonexpansive. Then for any sequences $(x_t)_{t \geqslant 1}$ in X,

$$\sum_{t=1}^{T} \|F(x_t) - x_t\|_2^2 \leqslant \frac{2}{\gamma} \underbrace{\sum_{t=1}^{T} \langle F(x_t) - x_t, x_* - x_t \rangle}_{\text{regret wrt } ((F(x_t) - x_t))_{t \geqslant 1}}, \qquad T \geqslant 1.$$

- No need to know γ to minimize the RHS.

AdaGrad

(McMahan–Streeter 2010)
(Duchi–Hazan–Singer, 2011)

- 3 main versions: AdaGrad-Norm, AdaGrad-Diagonal, AdaGrad-Full.
- A family of regret minimizing algorithms with adaptive guarantees.
- Time-dependent step-sizes based on previous data.
- Important breakthrough. Good theoretical properties and good behavior in practice.
- Lot of on-going research and new variants with improved properties.

AdaGrad-Norm: definition and regret bound

$$x_{t+1} = x_t + \frac{\eta}{\sqrt{\sum_{s=0}^{t} \|u_s\|_2^2}} u_t.$$

- Online gradient descent with adaptive step-size
  based on previously observed vectors
- Large vectors decrease subsequent step-sizes

Theorem (Regret bound for AdaGrad-Norm)

For all $T \geqslant 1$,

$$\sum_{t=1}^{T} \langle u_t, x_* - x_t \rangle \leqslant D_{\eta,T} \sqrt{\sum_{t=1}^{T} \|u_t\|_2^2}$$

$$\left(\textit{where} \quad D_{\eta,T} = \eta + \frac{\max_{1 \leqslant t \leqslant T} \|x_t - x_*\|_2^2}{2\eta} \right)$$

Adaptivity and robustness of AdaGrad-Norm in smooth convex optimization

$$\min_{x \in \mathbb{R}^d} f(x) \qquad \text{differentiable, } x_* \text{ a minimizer}$$

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{\sum_{s=0}^{t} \|\nabla f(x_s)\|_2^2}} \nabla f(x_t), \quad t \geqslant 1.$$

Theorem (Levy et al. 2018)

Let $L > 0$. If f is convex and ∇f is L-Lipschitz, for all $T \geqslant 1$,

$$\min_{1 \leqslant t \leqslant T} f(x_t) - f(x_*) \leqslant D_{\eta,T}^2 \frac{L}{T}.$$

- GD must choose step-size $1/L$, AdaGrad-Norm is adaptive to L.
- Local character of AdaGrad: L can be replaced by $L_T := \max_{1 \leqslant t \leqslant T} \frac{\|\nabla f(x_t)\|^2}{f(x_t) - f(x^*)}$.
- AdaGrad-Norm is also adaptive to the noise level in stochastic convex optimization

Nonexpansiveness and co-coercivity

Let $L > 0$.

Definition

An operator $G : X \to \mathbb{R}^n$ is L-co-coercive if for all $x, x' \in X$,

$$\langle G(x') - G(x), x' - x \rangle \geqslant \frac{1}{L} \|G(x') - G(x)\|_2^2 .$$

Proposition

Let $F : X \to X$ and $G = (I - F)/2$.

- *The fixed points of F are the zeros of G.*
- *F is nonexpansive iif G is 1-co-coercive,*
- *G is L-co-coercive iif $F_{1/L} = I - \frac{2}{L} G$ is nonexpansive.*

AdaGrad-Norm for fixed points: adaptive guarantee

$$x_{t+1} = x_t + \gamma \frac{F(x_t) - x_t}{\sqrt{\sum_{s=1}^{t} \|F(x_s) - x_s\|_2^2}}.$$

Theorem (K., 2024)

If $F_{1/L}$ is nonexpansive (i.e. $G = (I - F)/2$ is L-co-coercive),

$$\min_{1 \leqslant t \leqslant T} \|F(x_t) - x_t\|_2 \leqslant \frac{2D_{\gamma,T} L}{\sqrt{T}}, \qquad T \geqslant 1.$$

- Adaptive to L.
- Adaptivity is local: L by can replaced by

$$L_T := \underbrace{\sup_{1 \leqslant t \leqslant T} \frac{\|F(x_t) - x_t\|^2}{2\left\langle F(x_t) - x_t, x^* - x_t \right\rangle}}_{\text{local co-coercivity along trajectory wrt } x_*}.$$

On conditionning

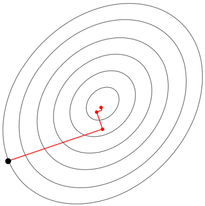
- For twice differentiable functions, optimality conditions at a minimizer give

$$\nabla f(x^*) = 0 \qquad \text{et} \qquad \nabla^2 f(x^*) \succeq 0.$$

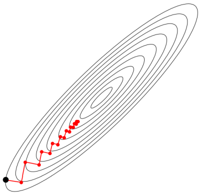
- As $x \to x^*$:

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^\top \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|^2).$$

- Local conditionning: $\kappa_{\text{loc}} := \frac{\lambda_{\max}(\nabla^2 f(x^*))}{\lambda_{\min}(\nabla^2 f(x^*))} \in [1, +\infty]$.



Small κ_{loc}: GD is fast
Well-conditionned case



Large κ_{loc}: GD is slow
Ill-conditionned case.

- A favorable change of coordinates $x \mapsto f(B^{-1}x)$ would improve conditionning

AdaGrad-Diagonal for optimization

$$x_{t+1} = x_t - \eta B_t^{-1} \nabla f(x_t)$$

$$\text{where} \quad B_t = \operatorname{diag}\left(\sqrt{\sum_{s=1}^{t}\left(\frac{\partial f}{\partial x_i}(x_s)\right)^2}\right)_{1 \leqslant i \leqslant d}$$

- Per-coordinate adaptive step-sizes.
- Partially addresses ill-conditionning

  by an online change of coordinates restricted to diagonal matrices.
- Much better scalability than quasi-Newton methods

  that maintain full matrices (thus needing $d \times d$ storage).
- Variants like RMSprop and Adam are state-of-the-art for DL.

  Some objective function varies much more/less wrt to some coordinates: weights of first vs last layers of a neural networks.

Generalized co-coercivity

Let $B \in \mathbb{R}^{d\times d}$ be symmetric positive definite.

Definition

An operator $G : X \to \mathbb{R}^d$ is co-coercive for B if for all $x, x' \in X$,

$$\langle G(x') - G(x), x' - x\rangle \geqslant \|G(x') - G(x)\|_{B^{-1}}^2 .$$

Proposition

G is co-coercive for B iif

$$I - 2B^{-1}G \quad \textit{is nonexpansive for } \|\cdot\|_B.$$

AdaGrad-Diagonal for fixed points: stronger adaptivity

$$x_{t+1} = x_t + \eta \left(\frac{(F(x_t) - x_t)_i}{\sqrt{\sum_{s=1}^{t} (F(x_s) - x_s)_i^2}} \right)_{1 \leqslant i \leqslant d}$$

Theorem (K., 2025)

Let $B \succ 0$ be a diagonal matrix. If $I - B^{-1}(F - I)$ is nonexpansive for $\|\cdot\|_B$,

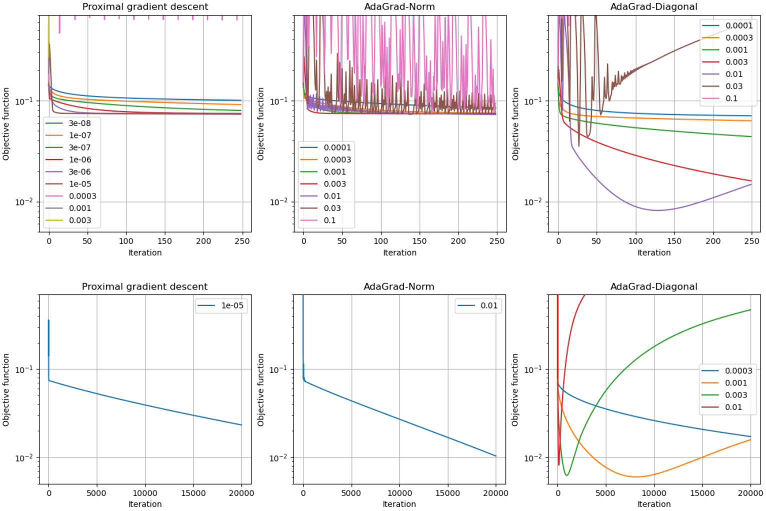
$$\min_{1 \leqslant t \leqslant T} \|F(x_t) - x_t\|_{B^{-1}} \leqslant D'_{\eta,T} \sqrt{\frac{\operatorname{Tr} B}{T}}.$$

$$\text{where} \quad D'_{\eta,T} = \frac{\max_{1 \leqslant t \leqslant T} \|x_t - x_*\|_\infty^2}{2\eta} + \eta$$

- Much stronger adaptivity: wrt all diagonal positive definite matrices.
  i.e. wrt the most favourable change of coordinates with diagonal matrices and not only wrt a scalar scaling
- Local character of adaptivity
  to be worked out

Numerical experiments: LASSO logistic regression with forward-backward splitting

$$\text{minimizer of } f(x) + \lambda \|x\|_1 \iff \text{fixed point of } \mathrm{Prox}_{\gamma\lambda\|\cdot\|_1}(x - \gamma\nabla f(x))$$



Questions and perspectives

- Additional adaptivity to contractive properties.
- Extension to stochastic approximations.

  Stochastic approximation correspond to Krasnoselskii-Mann iterations with noisy operator evalution. Interesting for reinforcement learning.
- Combine with Blackwell's approachability.

  Recent success in extensive form games (e.g. Poker) have been obtained with Blackwell-based regret minimizers. On bounded domains only.
- Combine with AdaGrad-Full to obtain quasi-Newton-like methods for fixed points.

  AdaGrad-Full maintain full matrices and offer even strong adaptivity. For problems of moderate size.
- Combine with successful AdaGrad variants e.g. RMSprop and Adam.

  RMSprop and Adam are variants of AdaGrad with weaker theoretical understanding but improved practical performance. Very sucessful in deep learning.

Thank you for your attention

Stefan Banach, *Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales*, Fundamenta mathematicae **3** (1922), no. 1, 133–181.

Jean-Pierre Baillon and Ronald E. Bruck, *The rate of asymptotic regularity is $O(1/\sqrt{n})$*, Lecture Notes in Pure and Applied Mathematics **178** (1996), 51–81.

Jean-Bernard Baillon and Georges Haddad, *Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones*, Israel Journal of Mathematics **26** (1977), 137–150.

Roberto Cominetti, José A Soto, and José Vaisman, *On the rate of convergence of Krasnoselskii–Mann iterations and their connection with sums of Bernoullis*, Israel Journal of Mathematics **199** (2014), no. 2, 757–772.

Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems **26** (2013).

Arthur P Dempster, Nan M Laird, and Donald B Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the royal statistical society: series B (methodological) **39** (1977), no. 1, 1–22.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm, *Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent*, Proceedings of the AAAI Conference on Artificial Intelligence **35** (2021), no. 6, 5363–5371.

James Hannan, *Approximation to Bayes risk in repeated play*, Contributions to the Theory of Games **3** (1957), 97–139.

Sergiu Hart and Andreu Mas-Colell, *A simple adaptive procedure leading to correlated equilibrium*, Econometrica **68** (2000), 1127–1150.

———, *A general class of adaptive strategies*, Journal of Economic Theory **98** (2001), no. 1, 26–54.

G.M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon **12** (1976), 747–756.

Mark Aleksandrovich Krasnosel'skii, *Two remarks on the method of successive approximations*, Uspekhi Matematicheskikh Nauk **10** (1955), no. 1, 123–127.

Yehuda Kfir Levy, Alp Yurtsever, and Volkan Cevher, *Online adaptive methods, universality and acceleration*, Advances in Neural Information Processing Systems, 2018, pp. 6500–6509.

Yurii Nesterov, *Primal-dual subgradient methods for convex problems*, Mathematical programming **120** (2009), no. 1, 221–259.

Arkadi Nemirovski and David B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley Interscience, 1983.

Lewis Fry Richardson, *The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam*, Philosophical Transactions of the Royal Society of London, Series A. **210** (1911), no. 459-470, 307–357.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione, *Regret minimization in games with incomplete information*, Advances in neural information processing systems **20** (2007), 1729–1736.