21 Extragradient

Goal

Extragradient, convergence and convergence rate, line search, single gradient/projection variants, forward-backward-forward

Alert 21.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 21.2: Problem

In this lecture we continue studying algorithms for solving a smooth variational inequality (VI):

$$\text{find } \mathbf{w}^* \in C \text{ such that } \forall \mathbf{w} \in C, \quad \langle \mathbf{w} - \mathbf{w}^*, \mathsf{T}\mathbf{w} \rangle \geq \langle \mathbf{w} - \mathbf{w}^*, \mathsf{T}\mathbf{w}^* \rangle \geq 0,$$

where $\mathsf{T} : C \subseteq \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous (not strongly) monotone.

We remind that with $\mathsf{T} = (\partial_{\mathbf{x}} f, \partial_{\mathbf{y}} \text{-} f)$ and $C = \mathsf{X} \times \mathsf{Y}$ we may reduce the minimax problem

$$\min_{\mathbf{x} \in \mathsf{X}} \max_{\mathbf{y} \in \mathsf{Y}} \; f(\mathbf{x}, \mathbf{y}) \qquad = \qquad \max_{\mathbf{y} \in \mathsf{Y}} \min_{\mathbf{x} \in \mathsf{X}} \; f(\mathbf{x}, \mathbf{y})$$

to the VI above (at least when a saddle point exists and f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$).

Algorithm 21.3: Extragradient (EG, Korpelevich 1976)

We have proved in Theorem 19.14 that GDA converges linearly if T is L-Lipschitz continuous and σ-strongly monotone. The latter assumption can be removed by regularization, as discussed in Remark 18.46 and Alert 19.5. Below we present yet another ingenious algorithm that removes the strongly monotone assumption and converges provably faster.

```
Algorithm: Extragradient for finding a zero of $\mathsf{T} = \mathsf{A} + \mathsf{B}$
  Input: $\mathbf{w}_0 \in \operatorname{dom} \mathsf{A} \subseteq \operatorname{dom} \mathsf{B}$
1 for $t = 0, 1, 2, \ldots$ do
2     choose step size $\eta_t > 0$
3     $\tilde{\mathbf{w}}_t = J_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t - \eta_t \mathsf{B}\mathbf{w}_t)$                // peek
4     $\mathbf{w}_{t+1} = J_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t - \eta_t \mathsf{B}\tilde{\mathbf{w}}_t)$      // update with peeked information
```

Korpelevich, G. M. (1976). "The extragradient method for finding saddle points and other problems". *Ekonomika i matematicheskie metody*, vol. 12, no. 4, pp. 747–756.

Theorem 21.4: Convergence of vanilla extra-gradient (EG)

Let $\mathsf{B} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *be* L*-Lipschitz and monotone,* $\mathsf{A} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *be maximal monotone with* $\operatorname{dom} \mathsf{A} \subseteq \operatorname{dom} \mathsf{B}$*, and the sum* $\mathsf{T} := \mathsf{A} + \mathsf{B}$ *be maximal monotone. Set* $\eta_t \in [0, 1/\mathsf{L}]$*. Then, the following estimate holds for the extra-gradient Line 4:*

$$\forall (\mathbf{w}, \mathbf{w}^*) \in \operatorname{gph} \mathsf{T}, \mathbf{a}^* \in \mathsf{A}\mathbf{w}, \quad \langle \tilde{\mathbf{z}}_t - \mathbf{w}, \mathbf{w}^* \rangle \leq \sum_{k=0}^{t} a_{t,k} \langle \tilde{\mathbf{w}}_k - \mathbf{w}, \mathsf{B}\tilde{\mathbf{w}}_k + \mathbf{a}^* \rangle \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{2H_t}, \quad \textit{where}$$

$$\tilde{\mathbf{z}}_t := \sum\nolimits_{k=0}^{t} a_{t,k}\tilde{\mathbf{w}}_k, \quad a_{t,k} := \eta_k/H_t, \qquad H_t := \sum\nolimits_{k=0}^{t} \eta_k.$$

- *If $H_t \to \infty$, then either $\mathsf{F} := \mathsf{T}^{-1}\mathbf{0} = \emptyset$ and $\|\tilde{\mathbf{z}}_t\| \to \infty$, or $\tilde{\mathbf{z}}_t \to \mathbf{z}_\infty \in \mathsf{F}$.*
- *If $0 < \liminf_t \eta_t \le \limsup_t \eta_t < 1/\mathsf{L}$ and assume $\mathsf{F} \neq \emptyset$, then $\mathbf{w}_t - \tilde{\mathbf{w}}_t \to 0$ and $\tilde{\mathbf{w}}_t \to \mathbf{w}_\infty \in \mathsf{F}$.*

Proof: The proof is similar to Theorem 19.3. Fix any $(\mathbf{w}, \mathbf{w}^*) \in \operatorname{gph} \mathsf{T}$ and $\mathbf{a}^* \in \mathsf{A}\mathbf{w}$. We apply firm nonexpansiveness of $J_{\mathsf{A}}^{\eta_t}$ (see Exercise 16.9) to $\mathbf{w}_{t+1}$ and to $\tilde{\mathbf{w}}_t$:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \|J_{\mathsf{A}}^{\eta_t}[\mathbf{w}_t - \eta_t \mathsf{B}\tilde{\mathbf{w}}_t] - J_{\mathsf{A}}^{\eta_t}(\mathbf{w} + \eta_t \mathbf{a}^*)\|_2^2 \\ &\le \|\mathbf{w}_t - \mathbf{w} - \eta_t(\mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^*)\|_2^2 - \|\mathbf{w}_t - \mathbf{w}_{t+1} - \eta_t(\mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^*)\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2 + 2\eta_t \left\langle \mathbf{w} - \mathbf{w}_{t+1}, \mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^* \right\rangle, \\ \|\tilde{\mathbf{w}}_t - \mathbf{w}\|_2^2 &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 + 2\eta_t \left\langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathsf{B}\mathbf{w}_t + \mathbf{a}^* \right\rangle. \end{aligned} \tag{21.1}$$

We set $\mathbf{w} = \mathbf{w}_{t+1}$ in (21.1) and add to the previous inequality:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 - \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 + 2\eta_t[\langle \tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}, \mathsf{B}\tilde{\mathbf{w}}_t - \mathsf{B}\mathbf{w}_t\rangle + \langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^*\rangle] \\ &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 - \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 + 2\eta_t[\|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2 \cdot \|\mathsf{B}\mathbf{w}_t - \mathsf{B}\tilde{\mathbf{w}}_t\|_2 + \langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^*\rangle] \\ \text{(}\mathsf{B}\ \mathsf{L}\text{-Lipschitz)} &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 - \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 + 2\eta_t[\|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2 \cdot \mathsf{L}\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2 + \langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^*\rangle] \\ &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 - (1 - \eta_t\mathsf{L})(\|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 + \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2) + 2\eta_t \langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathsf{B}\tilde{\mathbf{w}}_t + \mathbf{a}^*\rangle \qquad (21.2) \\ \text{(}\mathsf{B}\text{ monotone)} &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 - (1 - \eta_t\mathsf{L})(\|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 + \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2) + 2\eta_t \langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathbf{w}^*\rangle. \qquad (21.3) \end{aligned}$$

Since $\eta_t \mathsf{L} \le 1$ the middle term is negative. Divide both sides of (21.3) by $H_t := \sum_{k=0}^{t} \eta_k$ and sum from $k = 0$ to $k = t$:

$$\forall (\mathbf{w}, \mathbf{w}^*) \in \operatorname{dom} \mathsf{T}, \quad \langle \tilde{\mathbf{z}}_t - \mathbf{w}, \mathbf{w}^*\rangle \le \sum_{k=0}^{t} a_{t,k} \langle \tilde{\mathbf{w}}_k - \mathbf{w}, \mathsf{B}\tilde{\mathbf{w}}_k + \mathbf{a}^*\rangle \le \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{2H_t}.$$

When $H_t \to \infty$, it follows from the maximal monotonicity of T that any limit point of $\tilde{\mathbf{z}}_t$ is a zero. Therefore, either $\tilde{\mathbf{z}}_t$ blows up or there exists $\mathbf{w}_\star \in \mathsf{F}$, whose existence we assume from now on. Continuing from (21.3) where we set $\mathbf{w} = \mathbf{w}_\star$ (so that we may choose $\mathbf{w}^* = \mathbf{0}$), it follows that $\{\mathbf{w}_t\}$ is Fejér monotone w.r.t. F. We can thus verify Proposition 16.2 as in Theorem 19.3 to conclude that $\tilde{\mathbf{z}}_t$ converges to some $\mathbf{z}_\infty \in \mathsf{F}$.

If $\limsup_t \eta_t < 1/\mathsf{L}$, it follows from (21.3) (with $\mathbf{w} = \mathbf{w}_\star$) that

$$\|\tilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 + \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 \to 0 \text{ hence also } \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2 \to 0. \tag{21.4}$$

We need only prove any limit point of the Fejér sequence $\{\mathbf{w}_t\}$, or equivalently $\{\tilde{\mathbf{w}}_t\}$, is a zero. Indeed, from (21.3) we have:

$$0 \leftarrow \langle \mathbf{w}_{t+1} - \mathbf{w} - \mathbf{w}_t + \mathbf{w}, \mathbf{w}_{t+1} - \mathbf{w} + \mathbf{w}_t - \mathbf{w}\rangle = \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \|\mathbf{w}_t - \mathbf{w}\|_2^2 \le 2\eta_t \langle \mathbf{w} - \tilde{\mathbf{w}}_t, \mathbf{w}^*\rangle.$$

Since $\liminf_t \eta_t > 0$, passing to the limit completes the proof. ■

The proof here is patterned after Theorem 19.3. When $\mathsf{F} \neq \emptyset$, we only need the following weaker monotonicity property of B (to derive (21.4)):

$$\forall \mathbf{w}_\star \in \mathsf{F}, \ \forall (\mathbf{w}, \mathbf{b}^*) \in \operatorname{gph} \mathsf{B}, \quad \langle \mathbf{w} - \mathbf{w}_\star, \mathbf{b}^*\rangle \ge 0,$$

and we just apply continuity of B in (21.2) to conclude that any limit point of $\tilde{\mathbf{w}}_t$ is a zero.

Remark 21.5: Comparison

Thus, for a Lipschitz continuous monotone operator B, we have managed to weaken the condition on η_t and proved convergence of the direct sequence $\mathbf{w}_t$ without averaging! Setting $\eta_t \equiv \eta$ yields $O(1/t)$ rate of convergence (for the averaged sequence $\tilde{\mathbf{z}}_t$), which is significantly faster than the $\tilde{O}(1/\sqrt{t})$ rate in Theorem 19.3 (with $\eta_t = 1/(\sqrt{t}\ln^{1+\epsilon} t)$ for any $\epsilon > 0$). On the other hand, Golowich et al. (2020) recently proved that the direct sequence $\mathbf{w}_t$ converges at $O(1/\sqrt{t})$ rate, which is tight and significantly worse than the averaged sequence!

Golowich, N., S. Pattathil, C. Daskalakis, and A. Ozdaglar (2020). "Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems". In: *Proceedings of Thirty Third Conference on Learning Theory*, pp. 1758–1784.

Remark 21.6: Line search (Khobotov 1987)

Inspecting the proof of Theorem 21.4 for where the L-Lipschitz continuity of B is used, we realize that we can and perhaps should perform the following line search (particularly when L is not known in advance):

$$\eta_t = \min\left\{\bar{\eta}, \gamma\frac{\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2}{\|\mathsf{B}\mathbf{w}_t - \mathsf{B}\tilde{\mathbf{w}}_t\|_2}\right\}, \quad \text{where} \quad \gamma \in (0, 1)$$

and $\bar{\eta}$ is a rough estimate of $1/\mathsf{L}$. Note however that $\tilde{\mathbf{w}}_t$ itself depends on η_t, so we resort to line search in the spirit of Amijo:

```
$\eta_t \leftarrow 2\bar{\eta}$
$\mathbf{w}_t^* \leftarrow \mathsf{B}\mathbf{w}_t$
repeat
    $\eta_t \leftarrow \eta_t/2$
    $\tilde{\mathbf{w}}_t \leftarrow J_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t - \eta_t\mathbf{w}_t^*)$
until $\eta_t \leq \gamma\frac{\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2}{\|\mathbf{w}_t^* - \mathsf{B}\tilde{\mathbf{w}}_t\|_2}$
```

It is clear from our proof of Theorem 21.4 that the sequence $\{\mathbf{w}_t\}$ remains to be Fejér monotone w.r.t. the solution set hence $\{\mathbf{w}_t\}$ and also $\{\tilde{\mathbf{w}}_t\}$ are bounded. Thus, we only need B to be locally Lipschitz continuous, from which we immediately deduce that $\eta_t \geq \underline{\eta} > 0$ hence Theorem 21.4 continues to hold and line search runs for at most $1 + \ln(\bar{\eta}/\underline{\eta})$ iterations.

Khobotov, E. N. (1987). "Modification of the extra-gradient method for solving variational inequalities and certain optimization problems". *USSR Computational Mathematics and Mathematical Physics*, vol. 27, no. 5, pp. 120–127.

Remark 21.7: When not to use EG

Needless to say, we could apply the EG Line 4 to minimize any (smooth, convex) function. However, there is no advantage in doing so, since we get essentially the same convergence rate as gradient descent (GD) while doubling per-step cost and suffering a smaller step size (recall in EG $\eta_t \in (0, 1/L)$ while in GD $\eta_t \in (0, 2/L)$).

Similarly, there is no advantage of applying EG to finding a fixed point of a nonexpansion T. Although the naive iteration $\mathbf{w} \leftarrow \mathsf{T}(\mathbf{w})$ may not converge, Krasnosel'skiĭ's method (Krasnosel'skiĭ 1955; Schaefer 1957)

$$\mathbf{w} \leftarrow [(1-\eta)\mathrm{Id} + \eta\mathsf{T}]\mathbf{w}, \quad \textit{i.e.} \quad \text{the gradient algorithm } \mathbf{w} \leftarrow \mathbf{w} - \eta(\mathrm{Id} - \mathsf{T})\mathbf{w},$$

converges for any $\eta \in (0, 1)$ while EG can only allow $\eta \in (0, \frac{1}{2})$ (since $\mathrm{Id} - \mathsf{T}$ is 2-Lipschitz continuous).

Krasnosel'skiĭ, M. A. (1955). "Two remarks on the method of successive approximations". *Uspekhi Mat. Nauk*, vol. 10, no. 1, pp. 123–127.

Schaefer, H. (1957). "Über die Methode sukzessiver Approximationen". *Jahresbericht der Deutschen Mathematiker-Vereinigung*, vol. 59, no. 1, pp. 131–140.

Example 21.8: Lagrangian for minimax

Consider the following minimax problem:

$$\min_{\mathbf{x}\in\mathsf{X},\mathbf{g}(\mathbf{x})\leq\mathbf{0}} \max_{\mathbf{y}\in\mathsf{Y},\mathbf{h}(\mathbf{y})\leq\mathbf{0}} f(\mathbf{x},\mathbf{y}),$$

where we have separated some complicated functional constraints from the feasible domain of $\mathbf{x}$ and $\mathbf{y}$, respectively. We introduce the Lagrangian:

$$\min_{\mathbf{x}\in\mathsf{X}} \max_{\mathbf{y}\in\mathsf{Y},\boldsymbol{\mu}\geq\mathbf{0}} \min_{\boldsymbol{\nu}\geq\mathbf{0}} f(\mathbf{x},\mathbf{y})+\boldsymbol{\mu}^\top\mathbf{g}(\mathbf{x})-\boldsymbol{\nu}^\top\mathbf{h}(\mathbf{y}),$$

Under certain conditions on $\mathbf{y}$ we have strong duality w.r.t. $(\mathbf{y},\boldsymbol{\mu})$ and $\boldsymbol{\nu}$, so we may swap the inner max and min:

$$\min_{\mathbf{x}\in\mathsf{X},\boldsymbol{\nu}\geq\mathbf{0}} \max_{\mathbf{y}\in\mathsf{Y},\boldsymbol{\mu}\geq\mathbf{0}} f(\mathbf{x},\mathbf{y})+\boldsymbol{\mu}^\top\mathbf{g}(\mathbf{x})-\boldsymbol{\nu}^\top\mathbf{h}(\mathbf{y}).$$

Under further conditions on $\mathbf{x}$ we may have full strong duality to completely swap min and max:

$$\max_{\mathbf{y}\in\mathsf{Y},\boldsymbol{\mu}\geq\mathbf{0}} \min_{\mathbf{x}\in\mathsf{X},\boldsymbol{\nu}\geq\mathbf{0}} f(\mathbf{x},\mathbf{y})+\boldsymbol{\mu}^\top\mathbf{g}(\mathbf{x})-\boldsymbol{\nu}^\top\mathbf{h}(\mathbf{y}).$$

We may apply the EG Line 4 to solve the above primal-dual problems simultaneously (provided that all functions involved are continuously differentiable). This is where Khobotov's line search in Line 6 is convenient: verifying continuous differentiability and existence of a saddle point suffices, and the rest is left to line search, which can even accelerate convergence! This idea of course applies to any monotonic algorithm.

Example 21.9: EG for linear program

Following Korpelevich (1976), we apply EG to the linear program:

$$\mathfrak{p}_\star = \min_{\mathbf{u}\geq\mathbf{0}} \ \langle\mathbf{u},\mathbf{c}\rangle \qquad \text{s.t.} \quad A\mathbf{u}\geq\mathbf{b}$$

$$\mathfrak{d}^\star = \max_{\mathbf{v}\geq\mathbf{0}} \ \langle\mathbf{b},\mathbf{v}\rangle \qquad \text{s.t.} \quad A^\top\mathbf{v}\leq\mathbf{c}.$$

We assume the Lagrangian

$$\min_{\mathbf{u}\geq\mathbf{0}} \max_{\mathbf{v}\geq\mathbf{0}} \ \langle\mathbf{u},\mathbf{c}\rangle+\langle\mathbf{b}-A\mathbf{u},\mathbf{v}\rangle$$

has a unique saddle point $\mathbf{w}_\star=(\mathbf{u}_\star,\mathbf{v}^\star)$. Choose the step size η_t so that EG iterates $\mathbf{w}_t=(\mathbf{u}_t,\mathbf{v}_t)$ and $\tilde{\mathbf{w}}_t=(\tilde{\mathbf{u}}_t,\tilde{\mathbf{v}}_t)$ converge to a saddle point $\mathbf{w}_\star=(\mathbf{u}_\star,\mathbf{v}^\star)$. Let $J=\text{supp}(\mathbf{v}^\star)=\{j: v_j^\star\neq 0\}$ and $\bar{J}$ its complement. Similarly we define I and $\bar{I}$ for $\mathbf{u}_\star$.

Korpelevich, G. M. (1976). "The extragradient method for finding saddle points and other problems". *Ekonomika i matematicheskie metody*, vol. 12, no. 4, pp. 747–756.

Algorithm 21.10: Past extragradient (pEG, Popov 1980)

The following ingenious variant only requires 1 evaluation of the operator T but still 2 projections per step. Compared to the extragradient Line 4, we simply recycle the past evaluation $\mathsf{T}\tilde{\mathbf{w}}_{t-1}$ to replace $\mathsf{T}\mathbf{w}_t$, saving us 1 evaluation of T.

Algorithm: Past extragradient for solving a smooth monotone VI

```
Input: $\mathbf{w}_0 = \tilde{\mathbf{w}}_{-1} \in C \subseteq \operatorname{dom} \mathsf{T}$
for $t = 0, 1, 2, \ldots$ do
    choose step size $\eta_t > 0$
    $\tilde{\mathbf{w}}_t = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T} \tilde{\mathbf{w}}_{t-1})$
    $\mathbf{w}_{t+1} = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T} \tilde{\mathbf{w}}_t)$
```

Popov, L. D. (1980). “A modification of the Arrow-Hurwicz method for search of saddle points”. *Mathematical notes of the Academy of Sciences of the USSR*, vol. 28, no. 5, pp. 845–848.

Algorithm 21.11: Modified extragradient (mEG, Tseng 2000)

Another ingenious algorithm due to Tseng (2000) requires only 1 projection but still 2 evaluations of T per step. Note that this variant requires say $\operatorname{dom} \mathsf{T} = \mathbb{R}^d$.

Algorithm: Tseng’s modified forward-backward splitting

```
Input: $\mathbf{w}_0 \in C \subseteq \operatorname{dom} \mathsf{T}$
for $t = 0, 1, 2, \ldots$ do
    choose step size $\eta_t > 0$
    $\tilde{\mathbf{w}}_t = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T} \mathbf{w}_t)$
    $\mathbf{w}_{t+1} = \tilde{\mathbf{w}}_t - \eta_t(\mathsf{T} \tilde{\mathbf{w}}_t - \mathsf{T} \mathbf{w}_t)$
```

Tseng, P. (2000). “A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings”. *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 431–446.

Algorithm 21.12: Optimistic extragradient (oEG, Daskalakis et al. 2018)

Obviously, if we now combine the previous two ideas, we obtain a variant that only requires 1 projection and 1 evaluation of T per step!

Algorithm: Optimistic extragradient for solving a smooth monotone VI

```
Input: $\mathbf{w}_0 = \tilde{\mathbf{w}}_{-1} \in C \subseteq \operatorname{dom} \mathsf{T}$
for $t = 0, 1, 2, \ldots$ do
    choose step size $\eta_t > 0$
    $\tilde{\mathbf{w}}_t = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T} \tilde{\mathbf{w}}_{t-1})$
    $\mathbf{w}_{t+1} = \tilde{\mathbf{w}}_t - \eta_t(\mathsf{T} \tilde{\mathbf{w}}_t - \mathsf{T} \tilde{\mathbf{w}}_{t-1})$
```

Daskalakis, C., A. Ilyas, V. Syrgkanis, and H. Zeng (2018). “Training GANs with optimism”. In: *The 6th International Conference on Learning Representations*.

Algorithm 21.13: Reflected extragradient (rEG, Malitsky 2015)

Another variant that uses reflection and also enjoys 1 projection and 1 evaluation of T per step. Note that this variant requires say $\operatorname{dom} \mathsf{T} \supseteq 2C - C$.

Algorithm: Reflected extragradient for solving a smooth monotone VI

```
Input: $\mathbf{w}_0 = \mathbf{w}_{-1} \in C \subseteq \operatorname{dom} \mathsf{T}$
for $t = 0, 1, 2, \ldots$ do
    choose step size $\eta_t > 0$
    $\tilde{\mathbf{w}}_t = 2\mathbf{w}_t - \mathbf{w}_{t-1}$
    $\mathbf{w}_{t+1} = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T} \tilde{\mathbf{w}}_t)$
```

Malitsky, Y. (2015). "Projected Reflected Gradient Methods for Monotone Variational Inequalities". *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 502–520.

Remark 21.14: Mirror-Prox (Nemirovski 2004)

Nemirovski (2004) equipped the extragradient Line 4 with Bregman divergence and gave it a natural interpretation as approximation of the proximal point Algorithm 4.14. See also Nesterov (2007).

Nemirovski, A. (2004). "Prox-Method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems". *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251.

Nesterov, Y. (2007). "Dual extrapolation and its applications to solving variational inequalities and related problems". *Mathematical Programming*, vol. 109, no. 2, pp. 319–344.