# A Novel Catalyst Scheme for Stochastic Minimax Optimization

Yan Li

Georgia Institute of Technology

Joint work with George Lan

# Minimax Optimization and Proximal Point Method

# Minimax Optimization

## ▷ Problem of interest

$$\min_{x \in X} \left\{ f(x) := \max_{y \in Y} F(x, y) \right\}$$

- Many applications: machine learning (GAN, adversarial training); planning (robust MDPs, Markov games);
- Basic assumptions we will make:
  1. $\mu_p$ strongly convex in $x$; $\mu_d$ strongly concave in $y$;
  2. $\nabla F$ is $L$-Lipschitz
  3. We focus in this talk that $\mu_d, \mu_p > 0$, and WLOG $\mu_d \geq \mu_p$.

# A Brief Review of Existing Development

▷ **Many ways to solve this problem**

- Variational inequality (VI) based methods (Kotsalis et al., '20; Zhang et al., '23; many others): solve VI associated with the optimality condition

$$\langle \nabla_x F(x^*), x - x^* \rangle - \langle \nabla_y F(y^*), y - y^* \rangle \geq 0$$

1. Not optimal in the deterministic setting when $\mu_p \neq \mu_d$: $\mathcal{O}(L/\min\{\mu_p, \mu_d\}\log(1/\epsilon))$

# A Brief Review of Existing Development

▷ **Many ways to solve this problem**

- Variational inequality (VI) based methods (Kotsalis et al., '20; Zhang et al., '23; many others): solve VI associated with the optimality condition

$$\langle \nabla_x F(x^*), x - x^* \rangle - \langle \nabla_y F(y^*), y - y^* \rangle \geq 0$$

  1. Not optimal in the deterministic setting when $\mu_p \neq \mu_d$: $\mathcal{O}(L/\min\{\mu_p, \mu_d\}\log(1/\epsilon))$

- Primal-based methods: apply approximate proximal point framework
  1. (Near) Optimal complexity in the deterministic setting (Lin et al. '20):

$$\widetilde{\mathcal{O}}(L/\sqrt{\mu_p \mu_d}\log(1/\epsilon))$$

  2. Stochastic setting: seemingly no easy extension

# A Brief Review of Existing Development

▷ **Many ways to solve this problem**

- Variational inequality (VI) based methods (Kotsalis et al., '20; Zhang et al., '23; many others): solve VI associated with the optimality condition

$$\langle \nabla_x F(x^*), x - x^* \rangle - \langle \nabla_y F(y^*), y - y^* \rangle \geq 0$$

  1. Not optimal in the deterministic setting when $\mu_p \neq \mu_d$: $\mathcal{O}(L/\min\{\mu_p, \mu_d\}\log(1/\epsilon))$

- Primal-based methods: apply approximate proximal point framework
  1. (Near) Optimal complexity in the deterministic setting (Lin et al. '20):

$$\widetilde{\mathcal{O}}(L/\sqrt{\mu_p \mu_d}\log(1/\epsilon))$$

  2. Stochastic setting: seemingly no easy extension

- Can we design methods with optimal complexities in both deterministic and stochastic settings?

# Accelerated Proximal Point Method

**Algorithm** Accelerated Proximal Point Method

**Input:** initial points $\overline{x}_0 = \widetilde{x}_0$

**for** $k = 1, 2, \ldots, K$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1 - \gamma_k) x_{k-1}.$$

$$x_k = \min_{x \in X} f(x) + \frac{\beta_k}{2} \|x - \widehat{x}_k\|^2$$

$$\bar{x}_k = \frac{1}{\alpha_k \gamma_k + \mu(1-\gamma_k)} \left[ \beta_k x_k + (\mu_p - \beta_k)(1 - \gamma_k) x_{k-1} \right].$$

**end for**

**Output:** $\widetilde{x}_K$

- Convergence rate: with $\beta_k = \beta > 0$,

$$f(\widetilde{x}_K) - f(x^*) = \mathcal{O}\left(\frac{\beta}{K^2} \|x^* - \bar{x}_0\|^2\right)$$

# Accelerated Proximal Point Method

**Algorithm** Accelerated Proximal Point Method

**Input:** initial points $\overline{x}_0 = \widetilde{x}_0$

**for** $k = 1, 2, \ldots, K$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1 - \gamma_k) x_{k-1}.$$

$$x_k = \min_{x \in X} f(x) + \frac{\beta_k}{2} \|x - \widehat{x}_k\|^2$$

$$\bar{x}_k = \frac{1}{\alpha_k \gamma_k + \mu(1-\gamma_k)} \left[\beta_k x_k + (\mu_p - \beta_k)(1 - \gamma_k) x_{k-1}\right].$$

**end for**

**Output:** $\widetilde{x}_K$

- Convergence rate: with $\beta_k = \beta > 0$,

$$f(\widetilde{x}_K) - f(x^*) = \mathcal{O}(\frac{\beta}{K^2} \|x^* - \bar{x}_0\|^2)$$

- High-level idea of Catalyst: do approximate computation of proximal update using simple non-optimal methods

# Accelerated Proximal Point Method

**Algorithm** Accelerated Proximal Point Method

**Input:** initial points $\overline{x}_0 = \widetilde{x}_0$

**for** $k = 1, 2, \ldots, K$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1 - \gamma_k) x_{k-1}.$$

$$x_k = \min_{x \in X} f(x) + \frac{\beta_k}{2} \|x - \widehat{x}_k\|^2$$

$$\bar{x}_k = \frac{1}{\alpha_k \gamma_k + \mu(1-\gamma_k)} \left[\beta_k x_k + (\mu_p - \beta_k)(1 - \gamma_k) x_{k-1}\right].$$

**end for**

**Output:** $\widetilde{x}_K$

- Convergence rate: with $\beta_k = \beta > 0$,

$$f(\widetilde{x}_K) - f(x^*) = \mathcal{O}\left(\frac{\beta}{K^2} \|x^* - \bar{x}_0\|^2\right)$$

- Essential question: what error condition should we consider to measure "approximate computation"?

# Catalyst for Convex Optimization

# Catalyst for Convex Optimization

▷ **Problem of interest:** $\min_{x \in X} f(x)$, $f$ is $\mu$-strongly-convex and $L$-smooth

**Algorithm** `Catalyst`$(\mathcal{A})$: catalyst scheme for convex optimization

**Input:** initial points $\overline{x}_0 = \widetilde{x}_0$, to-be-catalyzed method $\mathcal{A}$.
**for** $k = 1, 2, \ldots, K$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1-\gamma_k)\widetilde{x}_{k-1}.$$

$$(\widetilde{x}_k, x_k) = \mathcal{A}(\phi_k, \widehat{x}_k)$$

$$\bar{x}_k = \frac{1}{\alpha_k \gamma_k + \mu(1-\gamma_k)} \left[\alpha_k x_k + (\mu - \alpha_k)(1-\gamma_k)\widetilde{x}_{k-1}\right].$$

**end for**
**Output:** $\widetilde{x}_K$

# Catalyst for Convex Optimization

▷ **Problem of interest:** $\min_{x \in X} f(x)$, $f$ is $\mu$-strongly-convex and $L$-smooth

**Algorithm** `Catalyst`$(\mathcal{A})$: catalyst scheme for convex optimization

**Input:** initial points $\bar{x}_0 = \widetilde{x}_0$, to-be-catalyzed method $\mathcal{A}$.
**for** $k = 1, 2, \ldots, K$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1-\gamma_k)\widetilde{x}_{k-1}.$$
$$(\widetilde{x}_k, x_k) = \mathcal{A}(\phi_k, \widehat{x}_k)$$
$$\bar{x}_k = \frac{1}{\alpha_k\gamma_k + \mu(1-\gamma_k)}\left[\alpha_k x_k + (\mu - \alpha_k)(1-\gamma_k)\widetilde{x}_{k-1}\right].$$

**end for**
**Output:** $\widetilde{x}_K$

1. $\phi_k$: subproblem of proximal step: $\phi_k(x) = f(x) + \frac{\beta_k}{2}\|x - \widehat{x}_k\|^2$
2. $\mathcal{A}(\phi_k, \widehat{x}_k)$: minimize $\phi_k$ using $\mathcal{A}$ (can be stochastic) initialized from $\widehat{x}_k$
3. Error condition:

$$\mathbb{E}[\phi_k(\widetilde{x}_k) - \phi_k(\widetilde{x}) + \frac{\alpha_k}{2}\|\widetilde{x} - x_k\|^2] \leq \frac{\varepsilon_k}{2}\|\widetilde{x} - \widehat{x}_k\|^2 + \delta_k, \ \forall \widetilde{x} \in X,$$

- $(\alpha_k, \epsilon_k)$ will depend both on $\beta_k$ and how long we run $\mathcal{A}$. Note that $\alpha_k$ is required for running Catalyst.

# Catalyst for Convex Optimization

▷ **Basic recursion**

$$\mathbb{E}[f(\widetilde{x}_k) - f(x) + \frac{\alpha_k \gamma_k^2 + \gamma_k(1-\gamma_k)\mu}{2}\|x - \bar{x}_k\|^2]$$
$$\leq (1-\gamma_k)[f(\widetilde{x}_{k-1}) - f(x)] + \frac{(\beta_k + \varepsilon_k)\gamma_k^2}{2}\|x - \bar{x}_{k-1}\|^2 + \delta_k.$$

## Lemma

*Suppose* $\mu = 0$. *Run* $\texttt{Catalyst}(\mathcal{A})$ *with*

$$\gamma_k = \frac{2}{k+1},\ \beta_k = \frac{(k+1)L}{k}.$$

*In addition, suppose* $\{\alpha_k\}$ *is chosen such that there exists* $\{(\varepsilon_k, \delta_k)\}$ *certifying error condition with*

$$\alpha_k = \beta_k(1+\varepsilon),\ \varepsilon_k = \beta_k \varepsilon,\ \varepsilon \leq 1,$$
$$\delta_k \leq \delta.$$

*Then we have*

$$\mathbb{E}[f(\widetilde{x}_K) - f(x^*)] \leq \frac{4L}{K^2}\|x^* - \bar{x}_0\|^2 + 2K\delta.$$

# Catalyzing SGD

▷ **What method to be catalyzed?** stochastic gradient descent

**Algorithm** $\texttt{SGD}(\phi_k; \widehat{x}_k)$: SGD for solving $\phi_k$, initialized at $\widehat{x}_k$

**Input:** stepsizes $\{\eta_t\}$, total number of steps $n > 0$
**for** t = 1, 2, . . . , T **do**
  Form $g_{t-1} = \nabla f(u_{t-1}; \xi_{t-1}) + \beta_k(u - \widehat{x}_k)$.
  $u_t = \operatorname{argmin}_{w \in X} \langle g_{t-1}, u \rangle + \frac{1}{2\eta_t} \|u - u_{t-1}\|^2$.
**end for**
Compute $\overline{u}_T$ = proper ergodic mean of $\{u_t\}$
**Output:** $(\overline{u}_T, u_T)$.

▷ $\texttt{SGD}$ satisfies error condition, with $(\alpha_k, \varepsilon_k)$ depending on $\beta_k$ and $T$.

**Proposition**

Let $(\widetilde{x}_k, x_k)$ be the output of $\texttt{SGD}(\phi_k; \widehat{x}_k)$, then

$$\mathbb{E}\left[\phi_k(\widetilde{x}_k) - \phi_k(u) + \frac{\alpha \mu_{\phi_k}}{2} \|\widetilde{u} - x_k\|^2\right] \leq \frac{\varepsilon \mu_{\phi_k}}{2} \|u - \widehat{x}_k\|^2 + \delta, \ \forall u \in X,$$

with $\varepsilon, \alpha, \delta$ depending on $T$, and $\mu_{\phi_k} = \mu + \beta_k$.

# Catalyzing SGD

**Theorem**

*Suppose $\mu = 0$. For any $\epsilon > 0$, run $\mathtt{Catalyst}$(SGD) with parameters*

$$K = 4\sqrt{\frac{L\|x^* - \overline{x}_0\|^2}{\epsilon}},\ \gamma_k = \frac{2}{k+1},\ \beta_k = \frac{(k+1)L}{k},\ \alpha_k = \frac{\beta_k}{1-\Lambda_T},$$

*where*

$$T = 8 + \frac{32\sigma^2 K}{L\epsilon},\ \Lambda_T = \frac{90}{(T+9)(T+10)}.$$

*At iteration $k$, the proximal step is approximately solved by running SGD$(\phi_k, \widehat{x}_k)$ for $T$ steps with stepsize*

$$\eta_t = \frac{2}{\beta_k(t+8)}.$$

*Then we have $\mathbb{E}[f(\widetilde{x}_K) - f(x^*)] \leq \epsilon$. The number of calls to SFO is bounded by*

$$\mathcal{O}\left(\sqrt{\frac{L\|x^* - \overline{x}_0\|^2}{\epsilon}} + \frac{\sigma^2\|x^* - \overline{x}_0\|^2}{\epsilon^2}\right).$$

# Catalyzing SGD

**Theorem**

*Suppose $\mu > 0$. Apply restarting strategy to the Catalyst framework. Then to obtain*

$$\mathbb{E}[f(\widetilde{x}_K) - f(x^*)] \leq \epsilon.$$

*The number of calls to SFO is bounded by*

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{f(x_{(0)}) - f(x^*)}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right).$$

- The complexities are optimal in both deterministic ($\sigma^2 = 0$) and stochastic ($\sigma^2 > 0$) settings, for both convex and strongly-convex objectives.

# Catalyst for Minimax Optimization

# Catalyst for Minimax Optimization

Denote $z = (x, y) \in X \times Y$.

**Algorithm** A minimax catalyst scheme

**Input:** initial points $\bar{z}_0 = \widetilde{z}_0 = z_0$

**for** $k = 1, 2, \ldots,$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1 - \gamma_k)\widetilde{x}_{k-1}.$$
$$(\widetilde{z}_k, z_k) = \mathcal{A}(\Phi_k, (\widehat{x}_k, y_{k-1}))$$
$$\bar{x}_k = \frac{1}{\alpha_k \gamma_k + \mu_p(1-\gamma_k)} \left[\alpha_k x_k + (\mu_p - \alpha_k)(1 - \gamma_k)\widetilde{x}_{k-1}\right].$$

**end for**

# Catalyst for Minimax Optimization

Denote $z = (x, y) \in X \times Y$.

**Algorithm** A minimax catalyst scheme

**Input:** initial points $\bar{z}_0 = \widetilde{z}_0 = z_0$
**for** $k = 1, 2, \ldots,$ **do**

$$\widehat{x}_k = \gamma_k \bar{x}_{k-1} + (1-\gamma_k)\widetilde{x}_{k-1}.$$

$$(\widetilde{z}_k, z_k) = \mathcal{A}(\Phi_k, (\widehat{x}_k, y_{k-1}))$$

$$\bar{x}_k = \frac{1}{\alpha_k\gamma_k + \mu_p(1-\gamma_k)}\left[\alpha_k x_k + (\mu_p - \alpha_k)(1-\gamma_k)\widetilde{x}_{k-1}\right].$$

**end for**

- The minimax Catalyst scheme looks almost identical to that of convex optimization
- $\mathcal{A}(\Phi_k, (\widehat{x}_k, y_{k-1}))$: $\min_{x\in X}\max_{y\in Y} \Phi_k(x, y) := F(x, y) + \frac{\beta_k}{2}\|x - \widehat{x}_k\|^2$ initialized at $(\widehat{x}_k, y_{k-1})$.
- Error condition:

$$\mathbb{E}\left[\Phi_k(\widetilde{x}_k, \widetilde{y}) - \Phi_k(\widetilde{x}, \widetilde{y}_k) + \frac{\alpha_k}{2}\|\widetilde{x} - x_k\|^2 + \frac{\alpha_k}{2}\|\widetilde{y} - y_k\|^2\right]$$

$$\leq \mathbb{E}\left[\frac{\varepsilon'_k}{2}\|\widetilde{x} - \widehat{x}_k\|^2 + \frac{\varepsilon_k}{2}\|\widetilde{y} - y_{k-1}\|^2\right] + \delta_k$$

# Catalyst for Minimax Optimization

## ▷ Basic recursion

$$\left(1-\frac{4\varepsilon_k}{\mu_d}\right)\mathbb{E}\left[f(\widetilde{x}_k)-f(x^*)\right]+\frac{\alpha_k\gamma_k^2+\gamma_k(1-\gamma_k)\mu_p}{2}\mathbb{E}\left[\|x^*-\bar{x}_k\|^2\right]+\frac{\alpha_k}{2}\mathbb{E}\left[\|\widetilde{y}_k^*-y_k\|\right]^2$$
$$\le\left(1-\gamma_k+\frac{4\varepsilon_k}{\mu_d}\right)\mathbb{E}\left[f(\widetilde{x}_{k-1})-f(x^*)\right]+\frac{(\beta_k+\varepsilon_k')\gamma_k^2}{2}\mathbb{E}\left[\|x^*-\bar{x}_{k-1}\|^2\right]$$
$$+\varepsilon_k\mathbb{E}\left[\|\widetilde{y}_{k-1}^*-y_{k-1}\|^2\right]+\delta_k.$$

# Catalyst for Minimax Optimization

▷ **Basic recursion**

$$\left(1-\frac{4\varepsilon_k}{\mu_d}\right)\mathbb{E}\left[f(\widetilde{x}_k)-f(x^*)\right]+\frac{\alpha_k\gamma_k^2+\gamma_k(1-\gamma_k)\mu_p}{2}\mathbb{E}\left[\|x^*-\bar{x}_k\|^2\right]+\frac{\alpha_k}{2}\mathbb{E}\left[\|\widetilde{y}_k^*-y_k\|\right]^2$$
$$\leq\left(1-\gamma_k+\frac{4\varepsilon_k}{\mu_d}\right)\mathbb{E}\left[f(\widetilde{x}_{k-1})-f(x^*)\right]+\frac{(\beta_k+\varepsilon_k')\gamma_k^2}{2}\mathbb{E}\left[\|x^*-\bar{x}_{k-1}\|^2\right]$$
$$+\varepsilon_k\mathbb{E}\left[\|\widetilde{y}_{k-1}^*-y_{k-1}\|^2\right]+\delta_k.$$

## Lemma

*Fix total iterations* $K\geq 1$ *a priori. Choose*

$$\gamma_k=\frac{2}{k+1},\ \beta_k=\frac{\mu_d(k+1)}{4(k+2)}.$$

*In addition, suppose* $\alpha_k$ *is chosen such that there exists* $\varepsilon_k$ *certifying error condition with*

$$\alpha_k=\beta_k(1+\varepsilon'),\ \varepsilon_k'=\beta_k\varepsilon',\ \varepsilon_k=\beta_k\varepsilon, \tag{3.1}$$
$$\delta_k\leq\delta, \tag{3.2}$$

*for some* $\delta>0$ *and* $\varepsilon\leq\min\left\{\frac{1}{12},\frac{1}{(K+1)(K+2)},\frac{\|x^*-\widetilde{x}_0\|^2}{2[f(\widetilde{x}_0)-f(x^*)]}\right\},\ \varepsilon'\leq 1$. *Then*

$$\mathbb{E}\left[f(\widetilde{x}_K)-f(x^*)\right]\leq\frac{24D_0\mu_d}{K^2}+64K\delta,\ D_0=\|x^*-\widetilde{x}_0\|^2+\|\widetilde{y}_0^*-y_0\|^2.$$

# Catalyst for Minimax Optimization

▷ **What methods do we catalyze?** The solution of the to-be-catalyzed method needs to certify the error condition.

# Catalyst for Minimax Optimization

▷ **What methods do we catalyze?** The solution of the to-be-catalyzed method needs to certify the error condition.

**Algorithm** $\texttt{SEG}(H; z_0)$: extragradient for $\min_{x\in X}\max_{y\in Y} H(x,y)$

**Input:** stepsizes $\{\eta_t\}$, total number of steps $n > 0$, initial point $z_0 \in Z$

**for** t = 0, 1, . . . , T-1 **do**

Define $G(z,\xi) = [\nabla_x H(z;\xi); -\nabla_y H(z,\xi)]$. Sample $\xi_t, \widehat{\xi}_t$, and update

$$\widehat{z}_t = \operatorname*{argmin}_{z\in Z} \eta_t \langle G(z_t,\xi_t), z\rangle + \tfrac{1}{2}\|z - z_t\|^2;$$

$$z_{t+1} = \operatorname*{argmin}_{z\in Z} \eta_t \left[\left\langle G(\widehat{z}_t,\widehat{\xi}_t), z\right\rangle + \tfrac{\mu}{2}\|z-\widehat{z}_t\|^2\right] + \tfrac{1}{2}\|z - z_t\|^2.$$

**end for**

Construct $\overline{z}_T$ = proper ergodic mean of $\{z_t\}$.

**Output:** $(\overline{z}_T, z_T)$.

# Catalyst for Minimax Optimization

## Lemma

*Suppose*

$$L\eta_t \leq 1/2,\ t \geq 0. \tag{3.3}$$

*Then for any $T \geq 1$, we have*

$$\mathbb{E}\left[F(\overline{x}_T, y) - F(x, \overline{y}_T) + \frac{\mu\Lambda_T}{2(\Lambda_T - \Lambda_0)}\|z - z_T\|^2\right]$$
$$\leq \frac{\mu\Lambda_0}{2(\Lambda_T - \Lambda_0)}\|z - z_0\|^2 + \frac{8\mu\sigma^2}{\Lambda_T - \Lambda_0}\sum_{t=0}^{T-1}\eta_t^2\Lambda_t,$$

*where*

$$\Lambda_t = \begin{cases} 1, & t = 0; \\ (1 + \mu\eta_{t-1})\Lambda_{t-1}, & t \geq 1. \end{cases}$$

# Catalyst for Minimax Optimization

**Theorem**

*Suppose* $\mu_p = 0, \mu_d > 0$. *For any* $\epsilon > 0$, *choose total number of iterations*

$$K \geq \sqrt{\frac{24\left[2\mu_d\|x^* - \widetilde{x}_0\|^2 + \mu_d\|\widetilde{y}_0^* - y_0\|^2\right]}{\epsilon}}.$$

*Choose* $\{(\gamma_k, \beta_k, \alpha_k)\}$ *properly. In addition, at the* $k$*-th iteration of the catalyst scheme, run SEG procedure for a total of* $T$ *steps with proper stepsizes* $\{\eta_t\}$ *and* $T$. *Then we obtain*

$$\mathbb{E}\left[f(\widetilde{x}_K) - f(x^*)\right] \leq \epsilon.$$

*The total number of calls to SFO can be bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{LD_0}{\sqrt{\mu_d \epsilon}} + \frac{\sigma^2 D_0}{\epsilon^2}\right),$$

*where* $D_0 = \|x^* - \widetilde{x}_0\|^2 + \|\widetilde{y}_0^* - y_0\|^2$.

# Catalyst for Minimax Optimization

**Theorem**

*Suppose $\mu_d \geq \mu_p > 0$. Apply the restarting strategy to the Catalyst scheme. Then to obtain*

$$\mathbb{E}\left[f(\widetilde{x}_{(e)}) - f(x^*)\right] \leq \epsilon. \tag{3.4}$$

*The total number of calls to SFO can be bounded by*

$$\widetilde{\mathcal{O}}\left(\tfrac{L}{\sqrt{\mu_d \mu_p}} \log_2\left(\tfrac{\Delta_0}{\epsilon}\right) + \tfrac{\sigma^2}{\mu_p \epsilon}\right),$$

*where* $\Delta_0 = f(\widetilde{x}_{(0)}) - f(x^*) + \frac{\mu_d}{12}\|\widetilde{y}^*_{(0)} - y_{(0)}\|^2$.

- Optimal complexities in both deterministic and stochastic setting.