ICML
International Conference
On Machine Learning

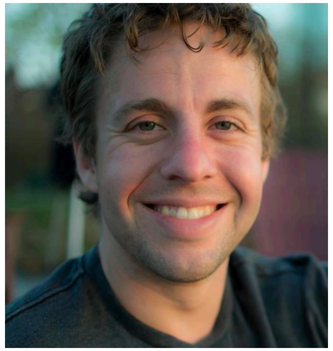
Stanford University

TEL AVIV UNIVERSITY

RECAPP: *Crafting a More Efficient Catalyst for Convex Optimization*

Yair Carmon, Arun Jambulapati, Yujia Jin, Aaron Sidford

(presenting)



Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization

- simple

  generalized from the “momentum” acceleration

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization

- simple

  generalized from the “momentum” acceleration

- applicable

  useful for lots of structured optimization through black-box reduction

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization

$$\min_x F(x)$$

- simple

  generalized from the “momentum” acceleration

- applicable

  useful for lots of structured optimization through black-box reduction

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization

$$\min_x F(x)$$

- simple

  generalized from the “momentum” acceleration

$$s_t \leftarrow (1 - \alpha_{t+1}) x_t + \alpha_{t+1} v_t$$

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2} \|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}} (s_t - x_{t+1})$$

- applicable

  useful for lots of structured optimization through black-box reduction

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization

$$\min_x F(x)$$

$$s_t \leftarrow (1-\alpha_{t+1})x_t + \alpha_{t+1} v_t$$

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - x_{t+1})$$

step-size independent of iterates

- simple

  generalized from the "momentum" acceleration

- applicable

  useful for lots of structured optimization through black-box reduction

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

- general acceleration framework for convex optimization
- simple

  generalized from the "momentum" acceleration
- applicable

  useful for lots of structured optimization through black-box reduction

$$\min_x F(x)$$

$$s_t \leftarrow (1-\alpha_{t+1})x_t + \alpha_{t+1}v_t$$

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - x_{t+1})$$

Reduce to regularized sub-problems

Usually require high-accuracy solutions

When accurate enough, converge at rate $O(\lambda R^2/T^2)$.

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).

[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.

[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.

[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

solve $\min_x F(x)$ to ϵ accuracy

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

solve $\min_x F(x)$ to ϵ accuracy

⬇

$O\left(\sqrt{\frac{\lambda}{\epsilon}} \cdot R\right)$ solves of subproblem $x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$

to sufficient high accuracy

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

Catalyst [LMH15]
aka. Approximate Proximal Point Acceleration (APPA) [G92, SS12, FGKS15]

solve $\min_x F(x)$ to ϵ accuracy

$O\left(\sqrt{\frac{\lambda}{\epsilon}} \cdot R\right)$ solves of subproblem $x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$

to sufficient high accuracy

MAIN Question: what accuracy is sufficient?

[LMH15] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "A universal catalyst for first-order optimization." *Advances in neural information processing systems* 28 (2015).
[G92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.
[SS12] Salzo, Saverio, and Silvia Villa. "Inexact and accelerated proximal point algorithms." *Journal of Convex analysis* 19.4 (2012): 1167-1192.
[FGKS15] Frostig, Roy, et al. "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization." *International Conference on Machine Learning*. PMLR, 2015.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

to sufficient high accuracy

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$ to sufficient high accuracy

Past answers in literature:

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \underbrace{\arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2}_{F^{\lambda}_{s_t}(x)}$$

to sufficient high accuracy

Past answers in literature:

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x \underbrace{F(x) + \frac{\lambda}{2}\|x - s_t\|^2}_{F_{s_t}^{\lambda}(x)}$$ to sufficient high accuracy

Past answers in literature:

[FGKS15] $F_s^{\lambda}(x) - F_s^{\lambda}(x_\star) \le O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^{\lambda}(x_{\text{init}}) - F_s^{\lambda}(x_\star)\right)$ for γ-strongly-convex F setting

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x \underbrace{F(x) + \frac{\lambda}{2}\|x - s_t\|^2}_{F_{s_t}^{\lambda}(x)}$$

to sufficient high accuracy

Past answers in literature:

- [FGKS15] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^\lambda(x_{\text{init}}) - F_s^\lambda(x_\star)\right)$ for γ-strongly-convex F setting
- [LMH18] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq \frac{\delta_t \lambda}{2}\|x - s\|^2$ where $\delta_t = (t+1)^{-2} \approx \epsilon / \lambda R^2$ for most iterations

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x \underbrace{F(x) + \frac{\lambda}{2}\|x - s_t\|^2}_{F^{\lambda}_{s_t}(x)}$$ to sufficient high accuracy

Past answers in literature:

- [FGKS15] $F_s^{\lambda}(x) - F_s^{\lambda}(x_\star) \le O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^{\lambda}(x_{\text{init}}) - F_s^{\lambda}(x_\star)\right)$ for γ-strongly-convex F setting
- [LMH18] $F_s^{\lambda}(x) - F_s^{\lambda}(x_\star) \le \frac{\delta_t \lambda}{2}\|x - s\|^2$ where $\delta_t = (t+1)^{-2} \approx \epsilon/\lambda R^2$ for most iterations
- [MS13] & [IPGSGM20] $\|\nabla F_s^{\lambda}(x)\| \le \sigma\lambda\|x - s\|$ for constant $\sigma \in (0,1)$

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_{x} \underbrace{F(x) + \frac{\lambda}{2}\|x - s_t\|^2}_{F^{\lambda}_{s_t}(x)}$$

to sufficient high accuracy

Past answers in literature:

- [FGKS15] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^\lambda(x_{\text{init}}) - F_s^\lambda(x_\star)\right)$ for γ-strongly-convex F setting
- [LMH18] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq \frac{\delta_t \lambda}{2}\|x - s\|^2$ where $\delta_t = (t+1)^{-2} \approx \epsilon / \lambda R^2$ for most iterations
- [MS13] & [IPGSGM20] $\|\nabla F_s^\lambda(x)\| \leq \sigma\lambda\|x - s\|$ for constant $\sigma \in (0,1)$

extraneous logarithmic factors for solving subproblems in *key applications*

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$ to sufficient high accuracy

Past answers in literature:

- [FGKS15] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^\lambda(x_{\text{init}}) - F_s^\lambda(x_\star)\right)$ for γ-strongly-convex F setting
- [LMH18] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq \frac{\delta_t \lambda}{2}\|x - s\|^2$ where $\delta_t = (t+1)^{-2} \approx \epsilon/\lambda R^2$ for most iterations
- [MS13] & [IPGSGM20] $\|\nabla F_s^\lambda(x)\| \leq \sigma\lambda\|x - s\|$ for constant $\sigma \in (0,1)$

Our proposal: Relaxed Error Condition for Accelerated Proximal Point (RECAPP)

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$ to sufficient high accuracy

Past answers in literature:

- [FGKS15] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^\lambda(x_{\text{init}}) - F_s^\lambda(x_\star)\right)$ for γ-strongly-convex F setting
- [LMH18] $F_s^\lambda(x) - F_s^\lambda(x_\star) \leq \frac{\delta_t \lambda}{2}\|x - s\|^2$ where $\delta_t = (t+1)^{-2} \approx \epsilon / \lambda R^2$ for most iterations
- [MS13] & [IPGSGM20] $\|\nabla F_s^\lambda(x)\| \leq \sigma\lambda\|x - s\|$ for constant $\sigma \in (0,1)$

Our proposal: Relaxed Error Condition for Accelerated Proximal Point (RECAPP)

$$F_s^\lambda(x) - F_s^\lambda(x_\star) \leq \frac{\frac{\lambda}{2}\|x_\star - s\|^2 + V_{x_\star}^F(x_t)}{8}$$

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

MAIN Question: what accuracy is sufficient?

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$ to sufficient high accuracy

Past answers in literature:

- [FGKS15] $F_s^\lambda(x) - F_s^\lambda(x_\star) \le O\left(\left(\frac{\gamma}{\lambda}\right)^{1.5}\right)\left(F_s^\lambda(x_{\text{init}}) - F_s^\lambda(x_\star)\right)$ for γ-strongly-convex F setting
- [LMH18] $F_s^\lambda(x) - F_s^\lambda(x_\star) \le \frac{\delta_t \lambda}{2}\|x - s\|^2$ where $\delta_t = (t+1)^{-2} \approx \epsilon/\lambda R^2$ for most iterations
- [MS13] & [IPGSGM20] $\|\nabla F_s^\lambda(x)\| \le \sigma\lambda\|x - s\|$ for constant $\sigma \in (0,1)$

Our proposal: Relaxed Error Condition for Accelerated Proximal Point (RECAPP)

$$F_s^\lambda(x) - F_s^\lambda(x_\star) \le \frac{\frac{\lambda}{2}\|x_\star - s\|^2 + V_{x_\star}^F(x_t)}{8}$$

Bregman divergence w.r.t previous iter x_t

[LMH18] Lin, Hongzhou, Julien Mairal, and Zaid Harchaoui. "Catalyst acceleration for first-order convex optimization: from theory to practice." *Journal of Machine Learning Research* 18.1 (2018): 7854-7907.
[MS13] Monteiro, Renato DC, and Benar Fux Svaiter. "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods." *SIAM Journal on Optimization* 23.2 (2013): 1092-1125.
[IPGSGM20] Ivanova, Anastasiya, et al. "Adaptive catalyst for smooth convex optimization." *International Conference on Optimization and Applications*. Springer, Cham, 2021.

RECAPP: full algorithm

$$s_t \leftarrow (1 - \alpha_{t+1})x_t + \alpha_{t+1}v_t$$

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - x_{t+1})$$

RECAPP: full algorithm

$$s_t \leftarrow (1-\alpha_{t+1})x_t + \alpha_{t+1}v_t$$

$$x_{t+1} \leftarrow \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - x_{t+1})$$

separating two x based on their different use in updates

RECAPP: full algorithm

$$s_t \leftarrow (1 - \alpha_{t+1}) x_t + \alpha_{t+1} v_t$$

$$x_{t+1} \overset{①}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - x_{t+1})$$

[BP15] Blanchet, Jose H., and Peter W. Glynn. "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization." *2015 Winter Simulation Conference (WSC)*. IEEE, 2015.
[ACJJS21] Asi, Hilal, et al. "Stochastic bias-reduced gradient methods." *Advances in Neural Information Processing Systems* 34 (2021): 10810-10822.

RECAPP: full algorithm

$$s_t \leftarrow (1-\alpha_{t+1})\boldsymbol{x_t} + \alpha_{t+1} v_t$$

$$x_{t+1} \overset{①}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$\tilde{x}_{t+1} \overset{②}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - \boldsymbol{\tilde{x}_{t+1}})$$

[BP15] Blanchet, Jose H., and Peter W. Glynn. "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization." *2015 Winter Simulation Conference (WSC)*. IEEE, 2015.
[ACJJS21] Asi, Hilal, et al. "Stochastic bias-reduced gradient methods." *Advances in Neural Information Processing Systems* 34 (2021): 10810-10822.

RECAPP: full algorithm

Relaxed Error Condition

$$s_t \leftarrow (1-\alpha_{t+1})x_t + \alpha_{t+1}v_t$$

$$x_{t+1} \overset{①}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$\tilde{x}_{t+1} \overset{②}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - \tilde{x}_{t+1})$$

① $$F^{\lambda}_{s_t}(x) - F^{\lambda}_{s_t}(x_\star) \le \frac{\frac{\lambda}{2}\|x_\star - s_t\|^2 + V^F_{x_\star}(x_t)}{8}$$

② Unbiased-ness: $\mathbb{E}x = x_\star$

$$\mathbb{E}\|x - x_\star\|^2 \le \frac{\frac{\lambda}{2}\|x_\star - s_t\|^2 + V^F_{x_\star}(x_t)}{4\lambda}$$

[BP15] Blanchet, Jose H., and Peter W. Glynn. "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization." *2015 Winter Simulation Conference (WSC)*. IEEE, 2015.
[ACJJS21] Asi, Hilal, et al. "Stochastic bias-reduced gradient methods." *Advances in Neural Information Processing Systems* 34 (2021): 10810-10822.

RECAPP: full algorithm

$$s_t \leftarrow (1-\alpha_{t+1})x_t + \alpha_{t+1} v_t$$

$$x_{t+1} \overset{①}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$\tilde{x}_{t+1} \overset{②}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - \tilde{x}_{t+1})$$

Relaxed Error Condition

(original condition)

① $F^\lambda_{s_t}(x) - F^\lambda_{s_t}(x_\star) \le \frac{\frac{\lambda}{2}\|x_\star - s_t\|^2 + V^F_{x_\star}(x_t)}{8}$

② Unbiased-ness: $\mathbb{E}x = x_\star$

$\mathbb{E}\|x - x_\star\|^2 \le \frac{\frac{\lambda}{2}\|x_\star - s_t\|^2 + V^F_{x_\star}(x_t)}{4\lambda}$

[BP15] Blanchet, Jose H., and Peter W. Glynn. "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization." *2015 Winter Simulation Conference (WSC)*. IEEE, 2015.
[ACJJS21] Asi, Hilal, et al. "Stochastic bias-reduced gradient methods." *Advances in Neural Information Processing Systems* 34 (2021): 10810-10822.

RECAPP: full algorithm

$$s_t \leftarrow (1-\alpha_{t+1})x_t + \alpha_{t+1}v_t$$

$$x_{t+1} \overset{①}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$\tilde{x}_{t+1} \overset{②}{\approx} \arg\min_x F(x) + \frac{\lambda}{2}\|x - s_t\|^2$$

$$v_{t+1} \leftarrow v_t - \frac{1}{\alpha_{t+1}}(s_t - \tilde{x}_{t+1})$$

Relaxed Error Condition

(original condition)

① $$F^\lambda_{s_t}(x) - F^\lambda_{s_t}(x_\star) \le \frac{\frac{\lambda}{2}\|x_\star - s_t\|^2 + V^F_{x_\star}(x_t)}{8}$$

reduction via multi-level Monte-Carlo [BP15, CJJS21]

② Unbiased-ness: $\mathbb{E}x = x_\star$

$$\mathbb{E}\|x - x_\star\|^2 \le \frac{\frac{\lambda}{2}\|x_\star - s_t\|^2 + V^F_{x_\star}(x_t)}{4\lambda}$$

[BP15] Blanchet, Jose H., and Peter W. Glynn. "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization." *2015 Winter Simulation Conference (WSC)*. IEEE, 2015.
[ACJJS21] Asi, Hilal, et al. "Stochastic bias-reduced gradient methods." *Advances in Neural Information Processing Systems* 34 (2021): 10810-10822.

Key observation

Relaxed Error Condition

$$F_s^\lambda(x) - F_s^\lambda(x_\star) \leq \frac{\frac{\lambda}{2}\|x_\star - s\|^2 + V_{x_\star}^F(x_t)}{8}$$

can be satisfied **without extraneous log** term when F is **finite-sum** or **min-the-max** structure

[JZ13] Johnson, Rie, and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction." *Advances in neural information processing systems* 26 (2013).
[N04] Nemirovski, Arkadi. "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* 15.1 (2004): 229-251.

Key observation

Relaxed Error Condition

$$F_s^\lambda(x) - F_s^\lambda(x_\star) \le \frac{\frac{\lambda}{2}\|x_\star - s\|^2 + V_{x_\star}^F(x_t)}{8}$$

can be satisfied **without extraneous log** term when F is **finite-sum** or **min-the-max** structure

finite-sum minimization

classical one epoch of SVRG [JZ14]

\+ initialize at s

\+ full gradient query at x_t

$n + L/\lambda$ **gradient queries suffice**

[JZ13] Johnson, Rie, and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction." *Advances in neural information processing systems* 26 (2013).
[N04] Nemirovski, Arkadi. "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* 15.1 (2004): 229-251.

Key observation

Relaxed Error Condition

$$F_s^\lambda(x) - F_s^\lambda(x_\star) \le \frac{\frac{\lambda}{2}\|x_\star - s\|^2 + V_{x_\star}^F(x_t)}{8}$$

can be satisfied **without extraneous log** term when F is **finite-sum** or **min-the-max** structure

finite-sum minimization

classical one epoch of SVRG [JZ14]
+ initialize at s
+ full gradient query at x_t

$n + L/\lambda$ **gradient queries suffice**

Max-structured minimization

classical mirror-prox [N04] on minimax objective
+ initialize x at s
+ initialize y at best response* of x_t

L/λ **gradient queries suffice**

[JZ13] Johnson, Rie, and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction." *Advances in neural information processing systems* 26 (2013).
[N04] Nemirovski, Arkadi. "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* 15.1 (2004): 229-251.

Conclusion and Results

RECAPP:

a *general* acceleration framework with relaxed error condition

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.

Conclusion and Results

RECAPP:

a *general* acceleration framework with <u>relaxed error condition</u>

finite-sum minimization

Gradient query complexity =

$$O\left(\sqrt{\frac{nLR^2}{\epsilon}} + n \log \log n\right)$$

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.

Conclusion and Results

RECAPP:

a *general* acceleration framework with relaxed error condition

finite-sum minimization

Gradient query complexity =

$$O\left(\sqrt{\frac{nLR^2}{\epsilon}} + n \log\log n\right)$$

- first under general acceleration framework
- matches best problem-algorithms [SYY20]

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.

Conclusion and Results

RECAPP:

a *general* acceleration framework with relaxed error condition

finite-sum minimization

Gradient query complexity =

$$O\left(\sqrt{\frac{nLR^2}{\epsilon}} + n \log\log n\right)$$

- first under general acceleration framework
- matches best problem-algorithms [SYY20]

Max-structured minimization

Gradient query complexity =

$$O\left(\frac{LR}{\sqrt{\mu\epsilon}}\right) + \widetilde{O}\left(\frac{L}{\mu} + \sqrt{\frac{LR^2}{\epsilon}}\right)$$

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.

Conclusion and Results

RECAPP:

a *general* acceleration framework with relaxed error condition

finite-sum minimization

Gradient query complexity =

$$O\left(\sqrt{\frac{nLR^2}{\epsilon}} + n \log\log n\right)$$

- first under general acceleration framework
- matches best problem-algorithms [SYY20]

Max-structured minimization

Gradient query complexity =

$$O\left(\frac{LR}{\sqrt{\mu\epsilon}}\right) + \widetilde{O}\left(\frac{L}{\mu} + \sqrt{\frac{LR^2}{\epsilon}}\right)$$

- improves over previous minimax complexity [YZKH20]
- obtains a weaker solution (no dual guarantee)

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.

Conclusion and Results

RECAPP:

a *general* acceleration framework with relaxed error condition

finite-sum minimization

Gradient query complexity =

$$O\left(\sqrt{\frac{nLR^2}{\epsilon}} + n\log\log n\right)$$

- first under general acceleration framework
- matches best problem-algorithms [SYY20]

Max-structured minimization

Gradient query complexity =

$$O\left(\frac{LR}{\sqrt{\mu\epsilon}}\right) + \widetilde{O}\left(\frac{L}{\mu} + \sqrt{\frac{LR^2}{\epsilon}}\right)$$

- improves over previous minimax complexity [YZKH20]
- obtains a weaker solution (no dual guarantee)

(strongly-convex setting + empirical performance of finite-sum algorithm also discussed in paper)

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.

Conclusion and Results

Thank You!

RECAPP:

a *general* acceleration framework with relaxed error condition

finite-sum minimization

Gradient query complexity =

$$O\left(\sqrt{\frac{nLR^2}{\epsilon}} + n \log\log n\right)$$

- first under general acceleration framework
- matches best problem-algorithms [SYY20]

Max-structured minimization

Gradient query complexity =

$$O\left(\frac{LR}{\sqrt{\mu\epsilon}}\right) + \widetilde{O}\left(\frac{L}{\mu} + \sqrt{\frac{LR^2}{\epsilon}}\right)$$

- improves over previous minimax complexity [YZKH20]
- obtains a weaker solution (no dual guarantee)

(strongly-convex setting + empirical performance of finite-sum algorithm also discussed in paper)

[SYY20] Song, Chaobing, Yong Jiang, and Yi Ma. "Variance reduction via accelerated dual averaging for finite-sum optimization." *Advances in Neural Information Processing Systems* 33 (2020).
[YZKH20] Yang, Junchi, et al. "A catalyst framework for minimax optimization." *Advances in Neural Information Processing Systems* 33 (2020): 5667-5678.