Robust and Risk-Averse Accelerated Gradient Methods

Mert Gürbüzbalaban

Rutgers University

Rutgers Business School Newark and New Brunswick

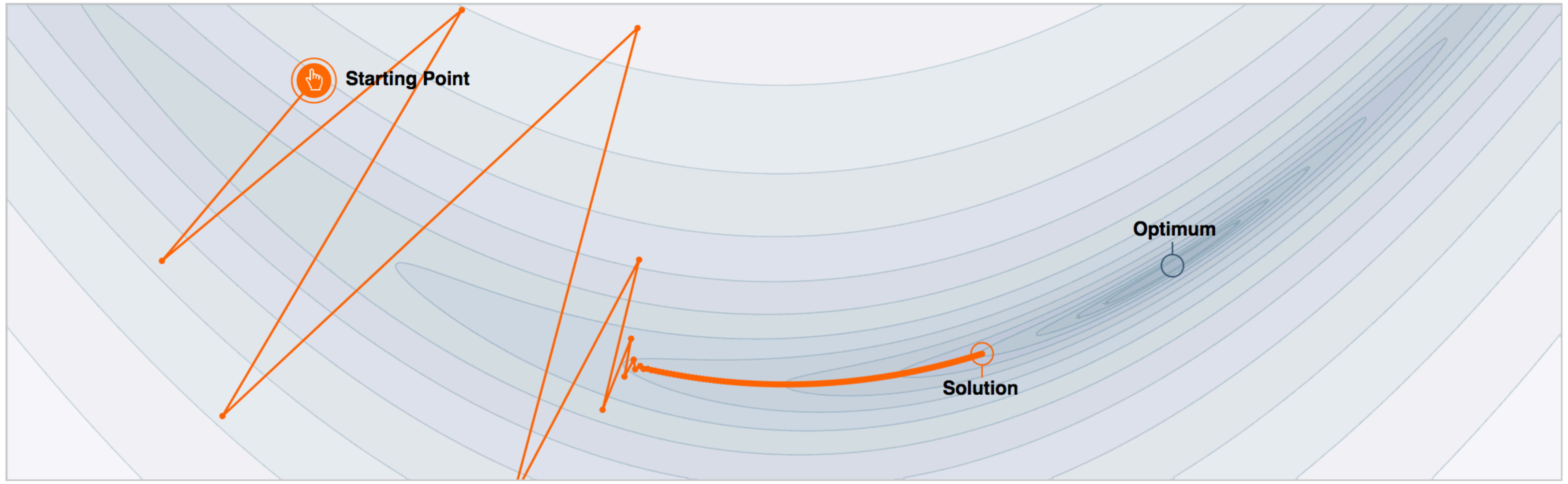
May 20th, 2022

Robustness and Resilience in Stochastic Optimization and Statistical Learning Workshop, Erice

First-Order Deterministic Optimization I

- Leading computational approach for large-scale optimization and machine learning.
- Simplest algorithm: **Gradient descent** (GD):

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$



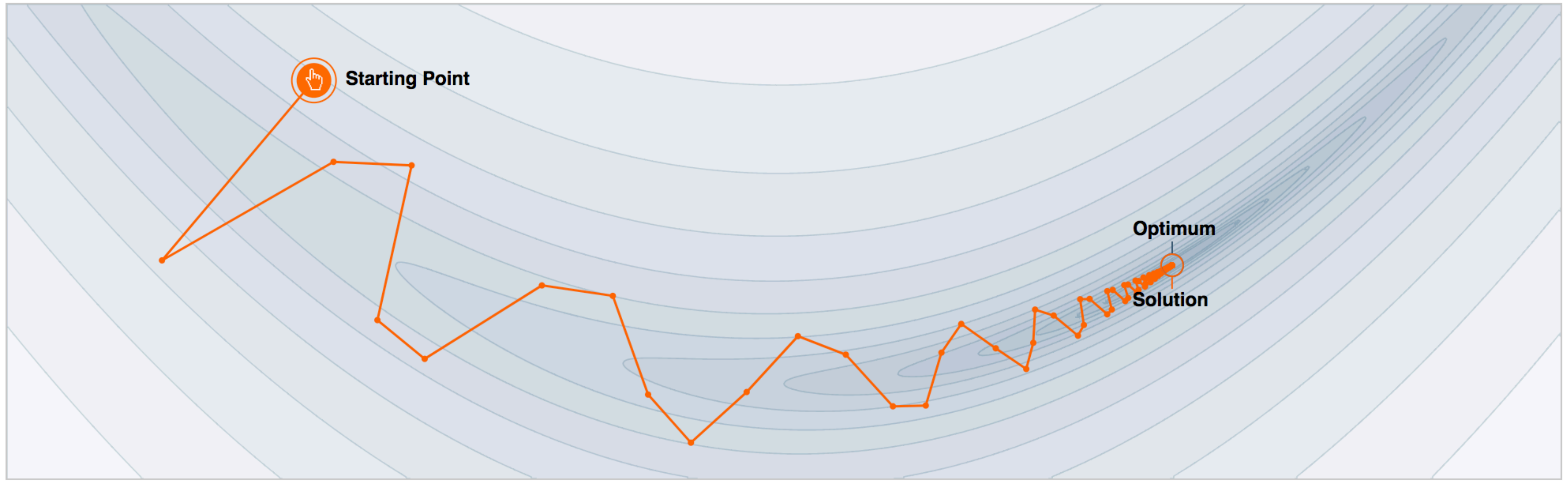
- When f is μ-strongly convex and L-smooth ($f \in \mathcal{S}_{\mu}^{L}(\mathbb{R}^d)$), linear rate ρ_{GD} achieved:

$$\alpha = \bar{\alpha} := \frac{2}{L+\mu} \implies \rho_{GD} = 1 - \frac{2}{\kappa + 1} \quad \text{with} \quad \kappa = \frac{L}{\mu}.$$

First-Order Deterministic Optimization II

- **Accelerated Gradient Descent** (AGD): [Nesterov, 1983]
  - ✦ Averages last two iterates for dampening oscillations.
  - ✦ Faster than gradient descent by tuning the momentum parameter β.

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(y_k),$$
$$y_{k+1} = x_k + \underbrace{\beta(x_k - x_{k-1})}_{\text{Momentum}},$$



- When f is μ-strongly convex and L-smooth ($f \in \mathcal{S}_\mu^L(\mathbb{R}^d)$), accelerated linear rate ρ_{acc}:

$$\alpha = \frac{1}{L}, \beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \implies \rho_{acc} = 1 - \frac{1}{\sqrt{\kappa}}.$$

More general $\alpha, \beta \implies$ rate $\rho(\alpha, \beta)$ [Hu, Lessard, ICML 2019]

Stochastic Optimization

- **In many settings, gradients contain random noise:**
  - Stochastic optimization or statistical learning setting: $f(x) = \mathbb{E}_{\omega \sim P} F(x, \omega)$
    - Example: Empirical risk minimization, logistic regression, linear regression.
  - Privacy-preserving empirical risk minimization.
- Consider

$$\min_{x \in \mathcal{X}} f(x),$$

where $\mathcal{X} \subset \mathbb{R}^d$ is compact and $f(x)$ is μ-strongly convex and L-smooth ($f \in \mathcal{S}_\mu^L(\mathbb{R}^d)$).

Assumption 1: We have only access to stochastic (noisy) estimate, $\tilde{\nabla} f(x)$, of the gradient $\nabla f(x)$, at the point $x \in \mathbb{R}^d$ satisfying

$$\mathbb{E}[\tilde{\nabla} f(x) - \nabla f(x) \,|\, x] = 0 \;\;\&\;\; \mathbb{E}[\|\tilde{\nabla} f(x) - \nabla f(x)\|^2 \mid x] \leq \sigma^2 \text{ for some } \sigma > 0. \qquad (L_p)$$

Triple momentum method (Generalized Momentum Methods)

- Unconstrained case ($\mathcal{X} = \mathbb{R}^d$)
- Triple momentum method (TMM):

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \tilde{\nabla} f(y_k),$$
$$y_{k+1} = x_k + \gamma(x_k - x_{k-1}),$$

More control !

- TMM is studied in [Hu & Lessard, 2017],[Scoy et al., 2018],[Cyrus et al., 2018] for deterministic optimization (fastest among deterministic first order algs.)
- TMM covers popular first order methods:
  - ✦ [$\gamma = \beta = 0$]: Gradient descent (GD),

$$x_{k+1} = x_k - \alpha \tilde{\nabla} f(y_k).$$

  - ✦ [$\gamma = 0$]: Heavy-ball method (HB),

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \tilde{\nabla} f(x_k).$$

  - ✦ [$\gamma = \beta$]: Nesterov's accelerated gradient descent (AGD),

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \tilde{\nabla} f(y_k),$$
$$y_{k+1} = x_k + \beta(x_k - x_{k-1}) .$$

Momentum: Sensivity to noise

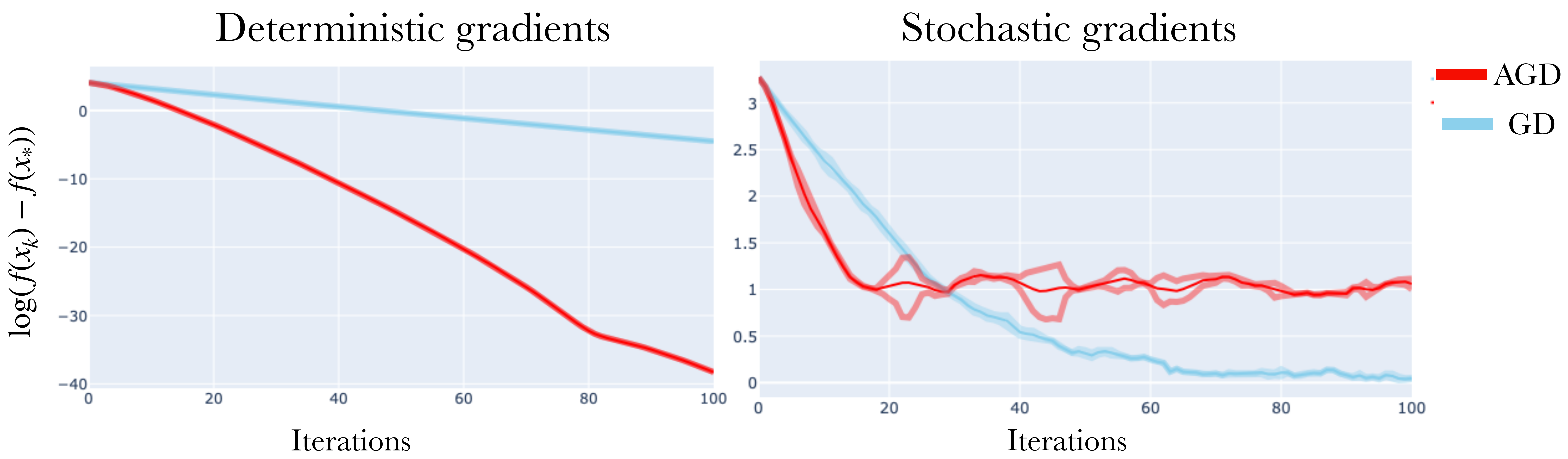


Figure: Standard AGD with $\alpha = 1/L$ and $\beta = (1 - \sqrt{1/\kappa})/(1 + \sqrt{1/\kappa})$ on quadratic objective under the various noise levels: $\sigma = 0$ (left) and $\sigma \gg 1$ (right)

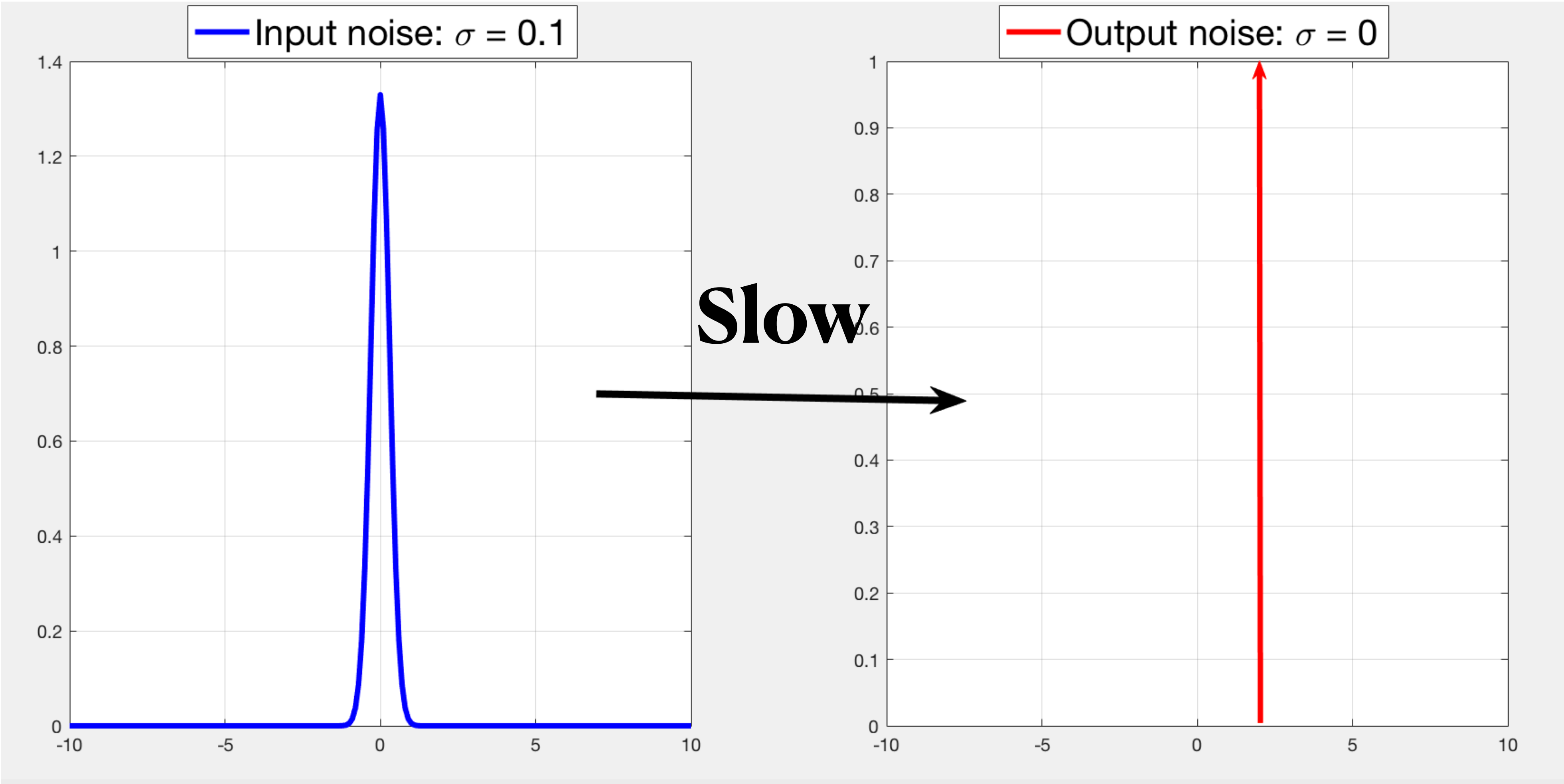
- Momentum methods are **sensitive** to persistent noise in the gradients [d'Aspremont, 2008], [Devolder, 2013], may even diverge[Flammiron & Bach, 2015].

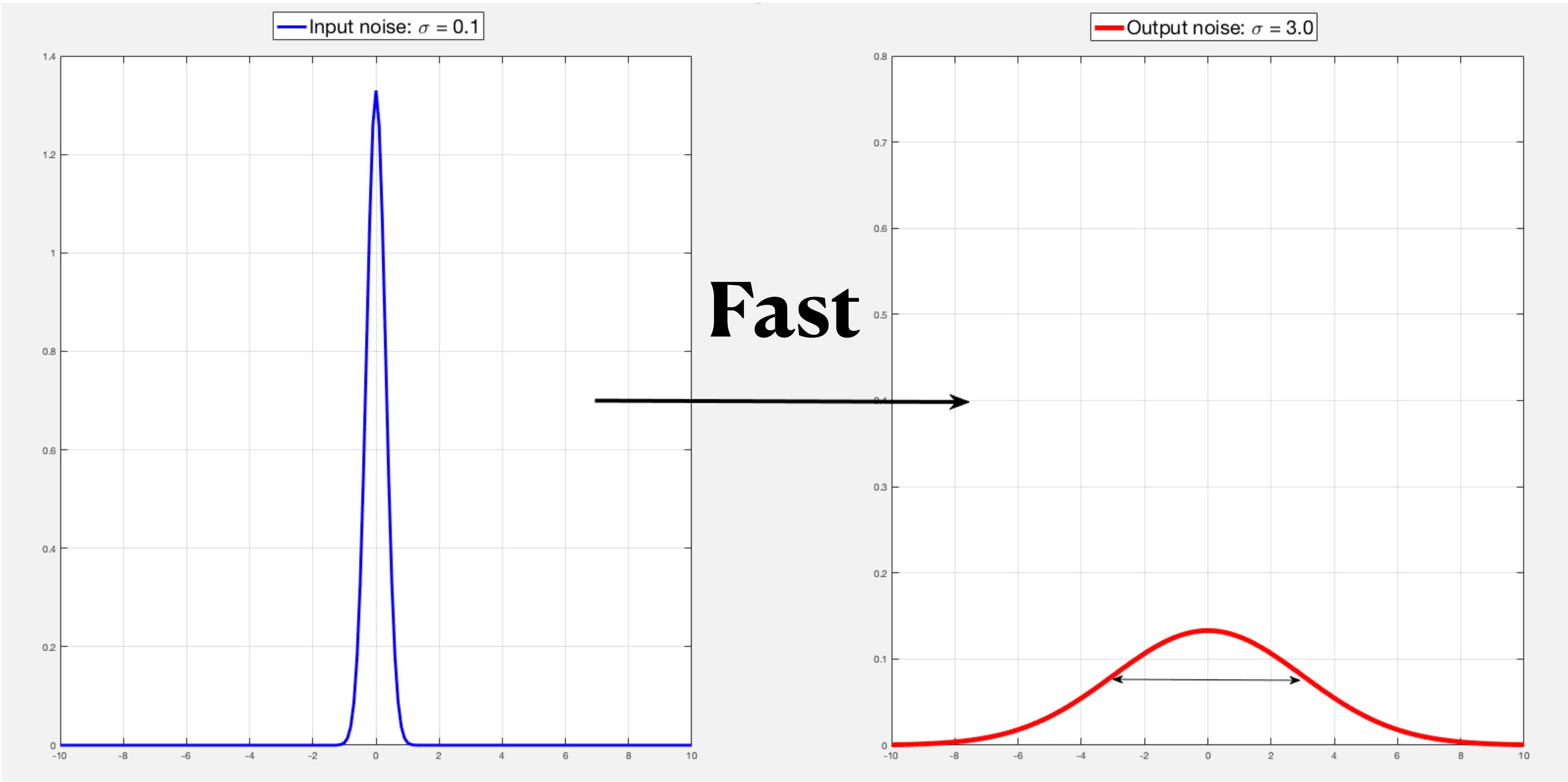
Rutgers Business School Newark and New Brunswick

Momentum: Effect of noise

- AGD with $\beta = \dfrac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$:

Noise Distribution | Stationary Distribution | Noise Distribution | Stationary Distribution





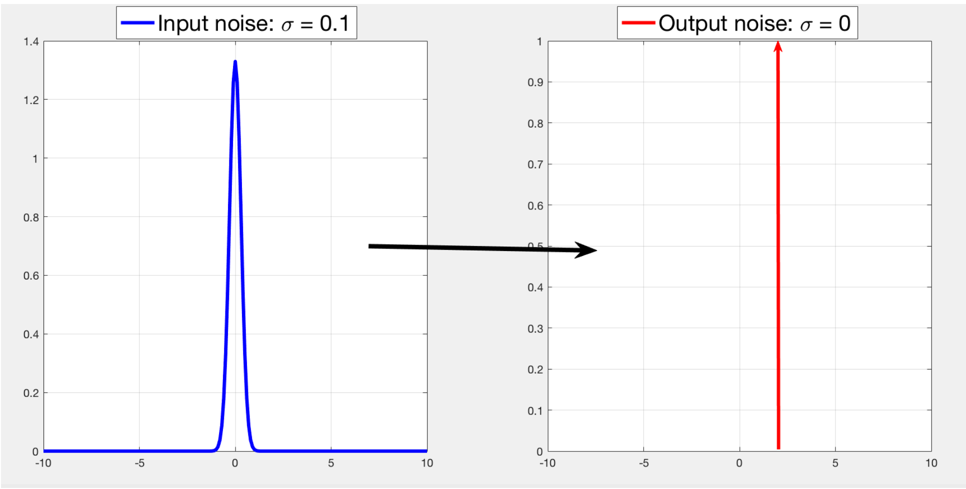
(a) $\alpha \approx 0$ **Slow but accurate (low variance/bias for suboptimality)**

(b) $\alpha = \dfrac{1}{L}$ **Fast but inaccurate (high variance/bias for suboptimality)**

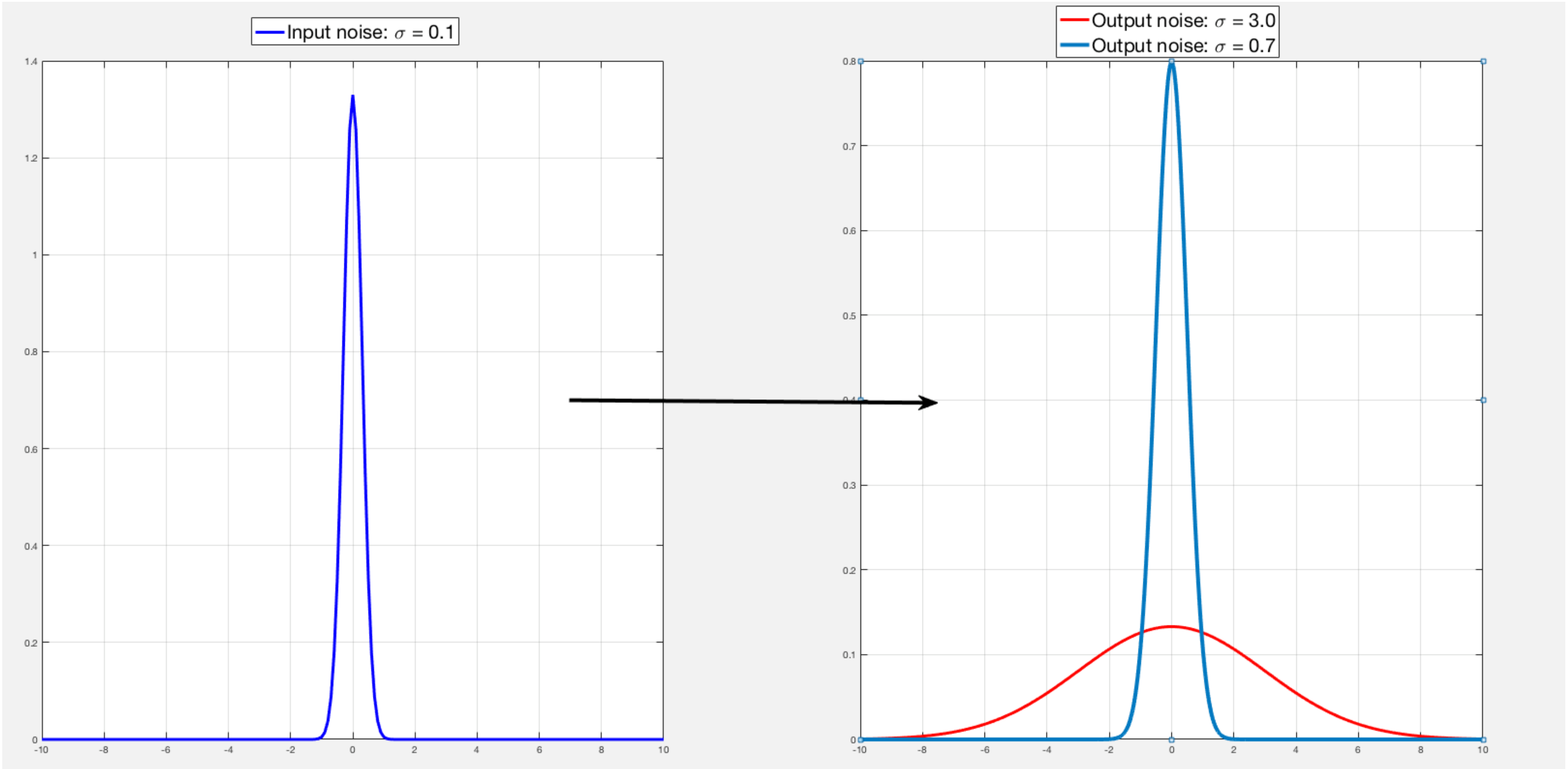
Rutgers Business School
Newark and New Brunswick

Momentum: Effect of noise

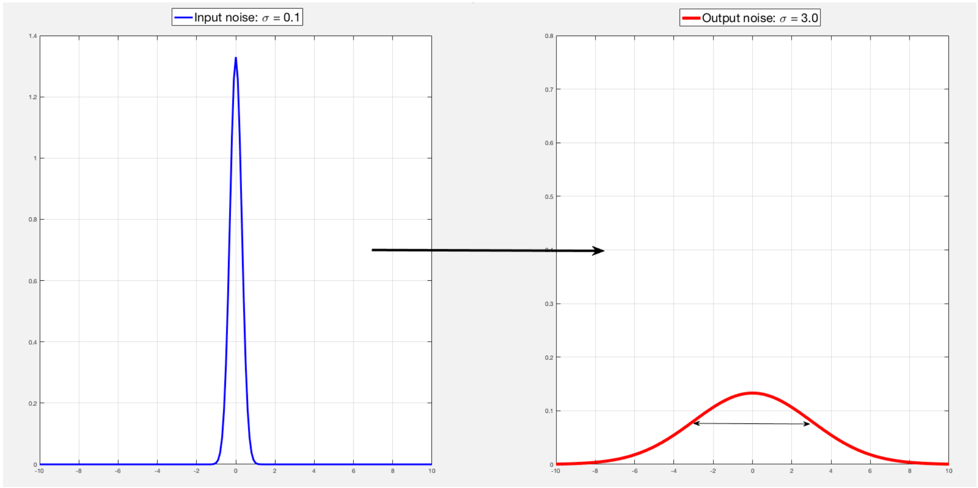
- **Input noise vs equilibrium distribution** for AGD with $\beta = \dfrac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$,



(a) $\alpha \approx 0$



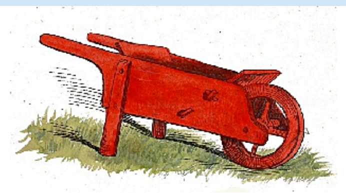
(c) $\alpha = \dfrac{0.1}{L}$

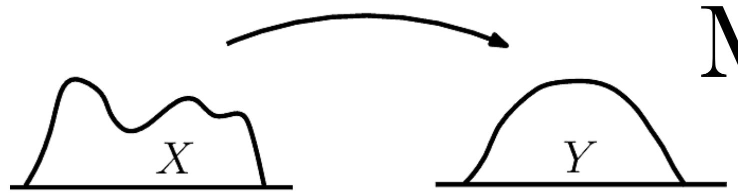


(b) $\alpha = \dfrac{1}{L}$

Theorem [Can, Zhu, M.G; ICML 2019]

Under some technical assumptions on the noise, the distribution π_k of AGD iterates $\{z_k\}$ converge linearly with rate $\rho(\alpha,\beta)$ w.r.t. 1-Wasserstein distance where $\rho(\alpha,\beta)$ is the rate of the (deterministic) accelerated GD algorithm.





Wasserstein distance btw X and Y:
Minimal cost of carrying sandpile X to sandpile Y

Re-usable proof technique for Bayesian learning with Langevin algorithms [**G.**, Gao, Hu. Zhu, JMLR 2021]

Momentum: Robustness to Noise

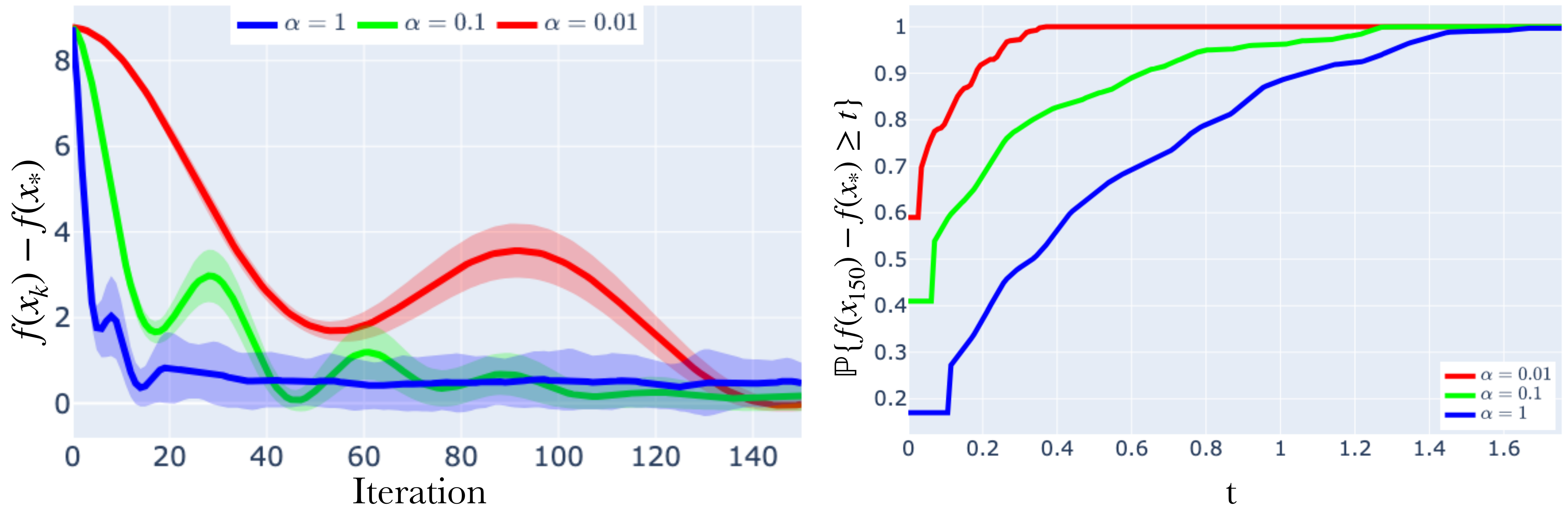


Figure: AGD algorithm with $\beta = (1 - \sqrt{\alpha\mu})/(1 + \sqrt{\alpha\mu})$ where the noise on the gradient is $\mathcal{N}(0, 16I_3)$ and the objective is quadratic function with $L = 10$ and $\mu = 0.01$. **Left:** The expected suboptimality and standard deviation from mean, **Right:** The histogram of $f(x_{150}) - f(x_*)$.

- **``Robustness to Noise"** /Noise Amplification: $\mathcal{J} := \limsup_{k \to \infty} \frac{1}{\sigma^2} \mathbb{E}[f(x_k) - f^*]$

  (BLUE HAS THE (WORST) LARGEST NOISE AMPLIFICATION.)

- Empirically: There is a trade-off between the convergence rate and robustness.

Heisenberg-like (Impossibility) Result

Proposition*

Let f be a quadratic with Hessian Q, for noisy GD with isotropic i.i.d. Gaussian noise we have:

$$\underbrace{\mathcal{J}(\alpha)}_{\text{noise amplification}} \cdot \underbrace{\frac{1}{1-\rho^2(\alpha)}}_{\text{convergence speed}} \geq c_f$$

for any choice of the stepsize for which $\rho(\alpha) < 1$ and $c_f := \frac{1}{8}\text{trace}(Q^{-2})$.

- **Faster convergence $\Longrightarrow$ worse** lower bound for **robustness.**
- Based on computing $\mathcal{J}(\alpha)$ and $\rho(\alpha)$ exactly for quadratics.
- Given rate, we can find the best parameters for optimizing robustness for strongly convex functions*

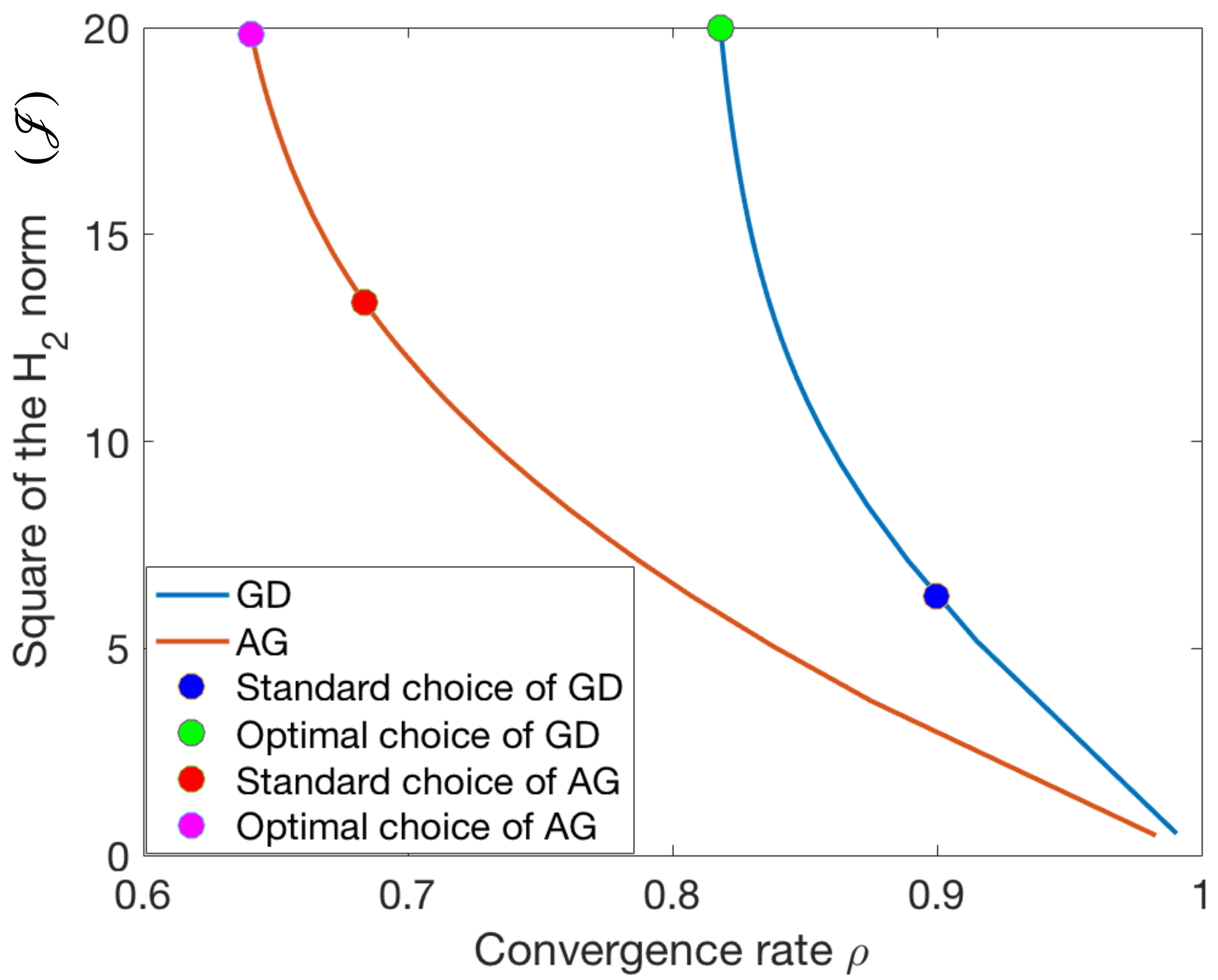


Fig: Best robustness achievable for given rate

* [Robust Accelerated Gradient Methods for Strongly Convex Functions, joint work with Aybat, Fallah, Ozdaglar SIOPT 2019]

Momentum: Effect on tail and the performance II

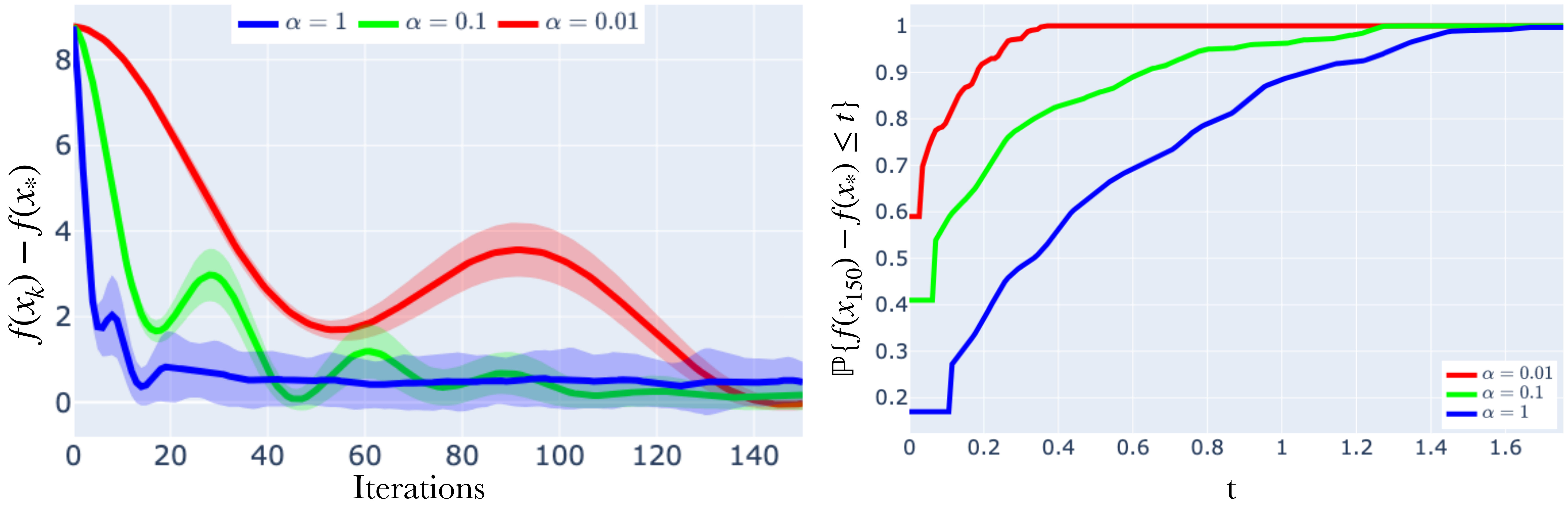


Figure: AGD algorithm with $\beta = (1 - \sqrt{\alpha\mu})/(1 + \sqrt{\alpha\mu})$ where the noise on the gradient is $\mathcal{N}(0,16I_3)$ and the objective is quadratic function with $L = 10$ and $\mu = 0.01$. **Left:** The expected suboptimality and standard deviation from mean, **Right:** The CDF of $f(x_{150}) - f(x_*)$.

- A stochastic dominance effect based on the choice of parameter.
- **The performance can be really bad unless the parameters are finely tuned!**
- **How to control the tail probabilities and deviation from mean as a function of parameters?**

Rutgers Business School
Newark and New Brunswick

Momentum: Effect on tail and the performance II

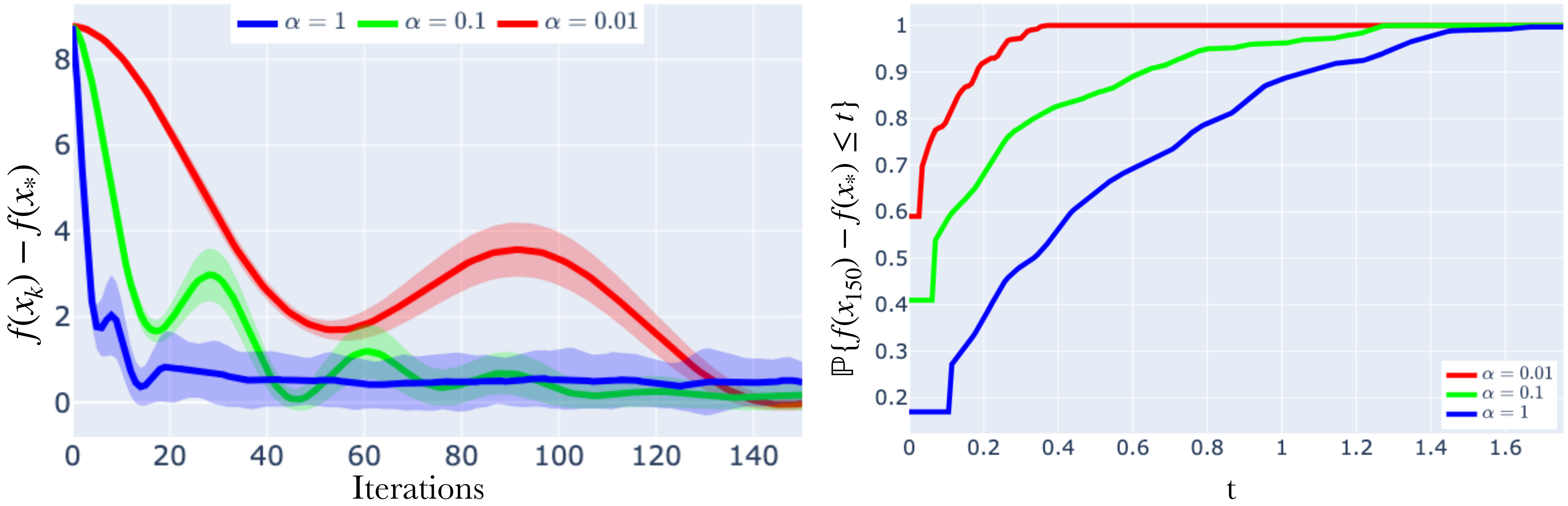


Figure: AGD algorithm with $\beta = (1 - \sqrt{\alpha\mu})/(1 + \sqrt{\alpha\mu})$ where the noise on the gradient is $\mathcal{N}(0,16I_3)$ and the objective is quadratic function with $L = 10$ and $\mu = 0.01$. **Left:** The expected suboptimality and standard deviation from mean, **Right:** The histogram of $f(x_{150}) - f(x_*)$.

- Next goal:
  - ✦ We want to understand the "risk", i.e. deviations from the mean.
  - ✦ The tail of the distribution π_k of the iterates $\{z_k\}$.

Entropic risk: Explaining tails

- **Finite-horizon entropic risk** at a given risk averseness $\theta > 0$ [Ruszczynski, 2013]:

$$r_{k,\sigma^2}(\theta) = \frac{2\sigma^2}{\theta} \log \mathbb{E}[e^{\frac{\theta}{2\sigma^2} f(x_k) - f(x_*)}],$$

- **Infinite-horizon entropic risk:**

$$r_{\sigma^2}(\theta) = \limsup_{k\to\infty} r_{k,\sigma^2}(\theta)$$

- Applying first-order Taylor expansion in θ:

$$r_{k,\sigma^2}(\theta) = \mathbb{E}[f(x_k) - f(x_*)] + \frac{\theta}{4\sigma^2}\mathbb{E}[\,|f(x_k) - f(x_*)|^2] + o(\theta).^\dagger$$

- The Chernoff bound:

$$\mathbb{P}\left\{ f(x_k) - f(x_*) \geq \boxed{\mathbf{r}_{\mathbf{k},\sigma^2}(\theta) + \frac{\mathbf{2\sigma^2}}{\theta}\log(\mathbf{1}/\zeta)} \right\} \leq \zeta,$$

where $\zeta \in (0,1)$ is the confidence level.

$\theta = 0$ (recovers the previous setting)

As $\theta \to 0$, risk measure converges to expected suboptimality

Bounds on the tail of suboptimality.

Entropic Risk controls quantiles

† See the paper for definition of little-o notation.

Entropic value at risk (EV@R): Coherent risk measure

- The **entropic value at risk** at a confidence level $\zeta \in (0,1)$ **:**

$$EV@R_{1-\zeta}[f(x_k) - f(x_*)] = \inf_{\theta > 0} \left\{ r_{k,\sigma^2}(\theta) + \frac{2\sigma^2}{\theta} \log(1/\zeta) \right\}.$$

- Smallest lower bound on tail:

$$\mathbb{P}\left(f(x_k) - f(x_*) \geq EV@R_{1-\zeta}[f(x_k) - f(x_*)]\right) \leq \zeta, \text{ for any } \zeta \in (0,1],$$

- Some properties of EV@R [Javid, 2012]:
  - ✦ A convex coherent risk measure,
  - ✦ The tightest possible upper bound obtained from Chernoff bound for the Value at Risk (V@R) of suboptimality,
  - ✦ An upper bound on the conditional value at risk (CV@R) of suboptimality.

Rutgers Business School
Newark and New Brunswick

Entropic value at risk (EV@R): Coherent risk measure

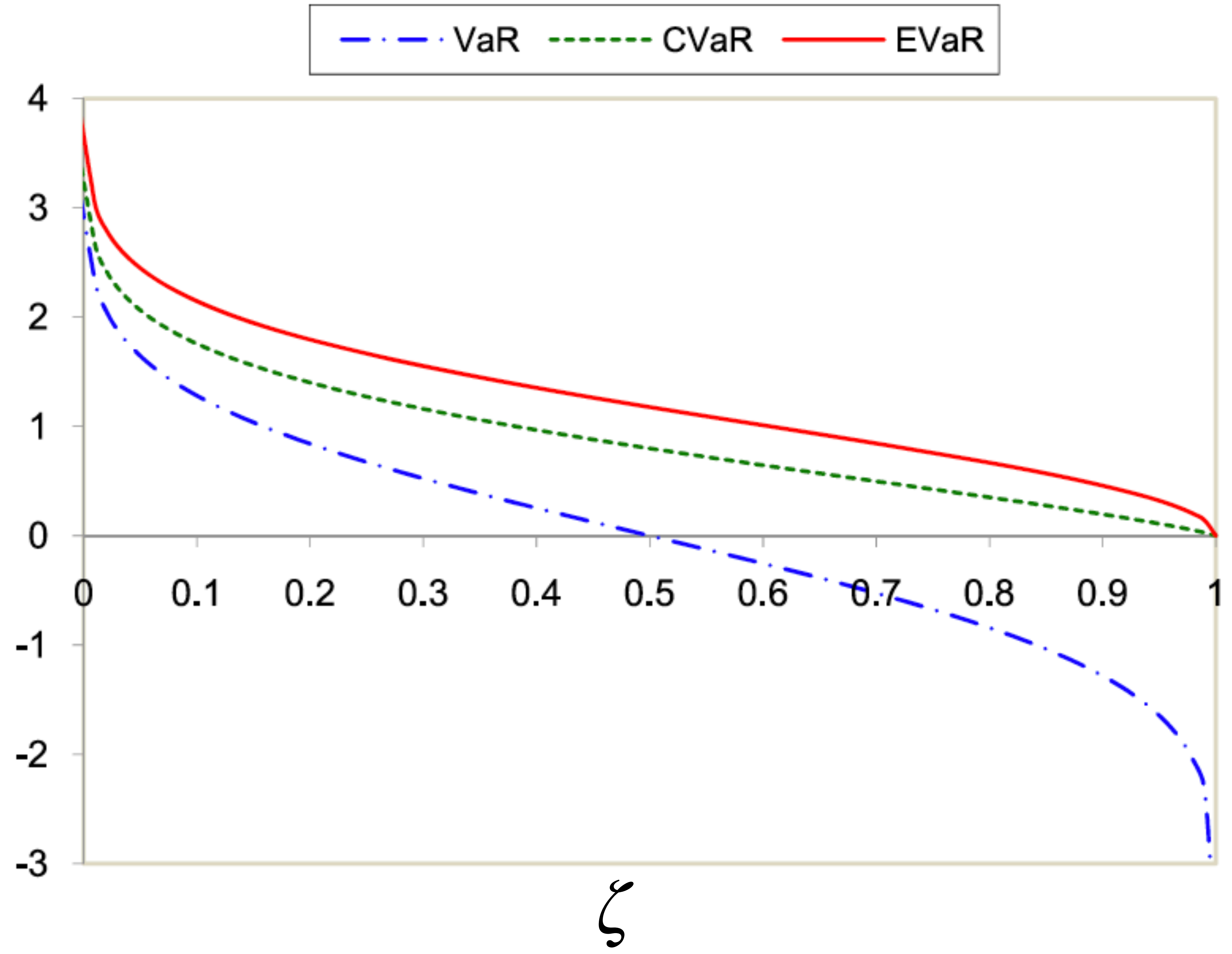
- The **entropic value at risk** at a confidence level $\zeta \in (0,1)$ **:**

$$EV@R_{1-\zeta}[f(x_k) - f(x_*)] = \inf_{\theta > 0} \left\{ r_{k,\sigma^2}(\theta) + \frac{2\sigma^2}{\theta} \log(1/\zeta) \right\}.$$

- Smallest lower bound on tail:

$$\mathbb{P}\left(f(x_k) - f(x_*) \geq EV@R_{1-\zeta}[f(x_k) - f(x_*)] \right) \leq \zeta, \text{ for any } \zeta \in (0,1],$$

- Some properties of EV@R [Javid, 2012]:
  - ✦ A coherent risk measure,
  - ✦ The tightest possible upper bound obtained from Chernoff bound for the Value at Risk (V@R) of suboptimality,
  - ✦ An upper bound on the conditional value at risk (CV@R) of suboptimality.



$EV@R_{1-\zeta}$,[X] $CV@R_{1-\zeta}$[X], and $V@R_{1-\zeta}$[X] comparison of standard normal distribution (taken from [Javid, 2018]).

Entropic value at risk (EV@R): Coherent risk measure

- The dual representation of EV@R

$$EV@R_{1-\zeta}[f(x_k) - f(x_*)] = \sup_{Q \in \mathcal{F}_{\pi_k}} \{\mathbb{E}_Q[f(x) - f(x_*)]\},$$

where $\mathcal{F}_{\pi_k} := \{Q \ll \pi_k \mid D_{KL}(Q||\pi_k) < \log(1/\zeta)\}$ and $D_{KL}(Q||\pi_k)$ is the KL divergence between Q and π_k.

- Interpreting duality:
  - EV@R is a robust version of expectation.
  - Worst-case expectation of $f(x_k) - f(x_*)$ w.r.t. measures around the $\log(1/\zeta)$ radius of π_k.

Our contributions*

- There are fundamental trade-offs between convergence rate and risk of suboptimality.
- Under some light tail assumption on the noise, for strongly convex optimization, we characterize the entropic risk of the suboptimality of TMM.
- We obtain finite-time performance bounds on the probability, $\mathbb{P}\{f(x_k) - f(x_*) \geq a\}$ for any $a > 0$ as a function of parameters.
- We study EV@R of the suboptimality which is a coherent risk measure capturing the deviations from the suboptimality.
- We propose a framework which systematically trade-offs the EV@R of suboptimality with the convergence rate to stationarity which allows us to obtain improved tail behavior for TMM.

*[Can, Gurbuzbalaban, Submitted, 2022].

TMM: Quadrative objectives

- Suppose f is convex quadratic with Hessian Q and w_{k+1} admits the **Assumption 2**[†]:

Assumption 2: For each $k \in \mathbb{N}$, $w_{k+1} = \tilde{\nabla} f(y_k) - \nabla f(y_k)$ is distributed according to isotropic Gaussian distribution, $\mathcal{N}(0, \sigma^2 I_d)$ for some $\sigma^2 > 0$, and it is independent from the filtration $\mathcal{F}_k$ generated by $\{x_j\}_{j=0}^{k}$.

Proposition 1

There exists $C_k = \mathcal{O}(k)$ we characterized explicitly such that

$$\|\mathbb{E}[z_k] - z_*\| \leq C_k \rho(A_Q)^{k-1} \|z_0 - z_*\|,$$

where $\rho(A_Q) := \max_{i \in \{1,..,d\}} \{\rho_i\}$ for $\rho_i = \begin{cases} \frac{1}{2}|c_i| + \frac{1}{2}\sqrt{c_i^2 + 4d_i}, & \text{if } c_i^2 + 4d_i > 0, \\ \sqrt{|d_i|}, & \text{otherwise,} \end{cases}$ for $c_i = (1+\beta) - \alpha(1+\gamma)\lambda_i(Q)$,

$d_i = -(\beta - \alpha\gamma\lambda_i(Q))$, and $\lambda_i(Q)$ is the i-th largest eigenvalue of the Hessian Q.

- Existing convergence results have been asymptotic [Gitman et al., 2019].

†We made the Assumption 2 for simplicity and our results can be extended to sub-Gaussian noise.

TMM: Quadrative objectives

- Suppose f is convex quadratic with Hessian Q and w_{k+1} admits the **Assumption 2,**

PROPOSITION 2

The finite-horizon risk measure is **finite** if and only if the parameters belong to

$$\mathcal{F}_\theta = \left\{ (\alpha, \beta, \gamma) \;\middle|\; |c_i| < |1 - d_i| \;\&\; \theta < 2 \min_{i \in \{1,..,d\}} \{u_i\}, \forall i \in \{1,..,d\} \right\}, \qquad (\theta\text{-feasible set})$$

where $u_i = \frac{(1 + d_i)[(1 - d_i)^2 - c_i^2]}{\lambda_i(Q)(1 - d_i)\alpha^2}$. Then (α, β, γ) also belong to

$$\mathcal{S}_q := \left\{ (\alpha, \beta, \gamma) \mid \rho(A_Q) < 1 \right\}, \qquad (\text{stable set})$$

and finite-horizon entropic risk linearly converges to infinite-horizon entropic risk, i.e.

$$|r_{k,\sigma^2}(\theta) - r_{\sigma^2}(\theta)| \leq \mathcal{O}(C_k^2 \rho(A_Q)^{2(k-1)} + C_k^4 \rho(A_Q)^{4(k-1)}) \qquad \text{for all } k \geq 1.^\dagger$$

† $\mathcal{O}(.)$ hides the constants depending on initialization.

Further discussion on $\mathscr{F}_\theta$ and $\mathscr{S}_q$

- For all $(\alpha, \beta, \gamma) \in \mathscr{F}_\theta$,

$$\|\mathbb{E}[x_k] - x_*\| \to 0,$$

- Particularly,

$$\mathscr{F}_\theta \subset \mathscr{S}_q,$$

- with the property that

$$\bigcup_{\theta>0} \mathscr{F}_\theta = \mathscr{S}_q.$$

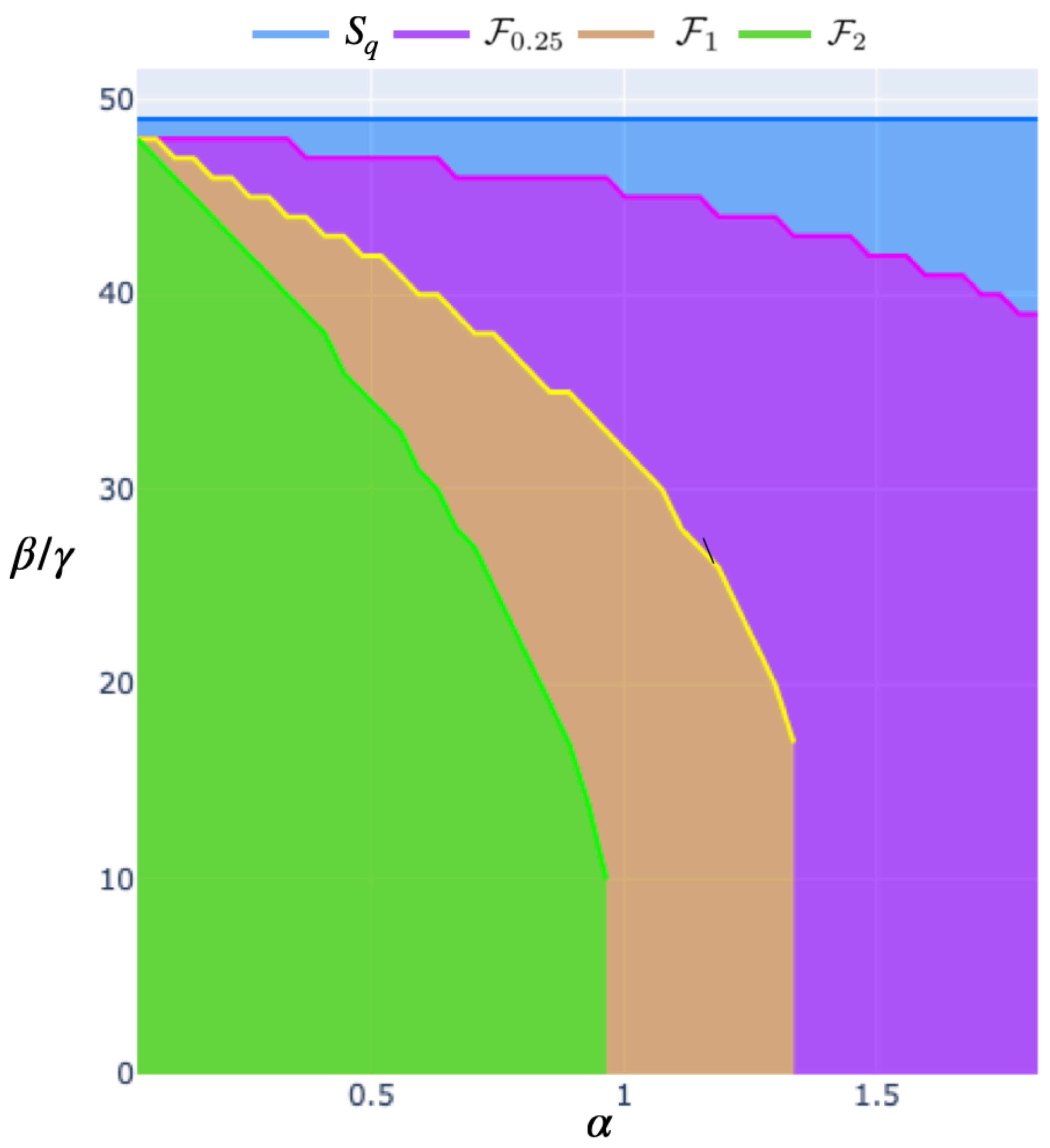


Figure: Feasible set vs stable set for $f(a, b) = a^2 + 0.1b^2$ where $a, b \in \mathbb{R}$ and $\sigma^2 = 1$.

EV@R of TMM on quadratic objectives

- Suppose f is convex quadratic with Hessian Q and w_{k+1} admits the **Assumption 2,**

Theorem 3

For $(\alpha, \beta, \gamma) \in \mathscr{F}_\theta$, we have $r_{\sigma^2}(\theta) = -\frac{\sigma^2}{\theta}\sum_{i=1}^{d}\log\left(1-\frac{\theta}{2u_i}\right)$.

Moreover let x_∞ be distributed according to stationary distribution of $\{x_k\}$, then

$$EV@R_{1-\zeta}[f(x_\infty)-f(x_*)] \le \bar{E}^q_{1-\zeta}(\alpha,\beta,\gamma) := \frac{\sigma^2}{\theta_0 2\bar{u}}\left[-d\log\left(1-\theta_0\right)+2\log(1/\zeta)\right]$$

$$\boxed{\le \frac{M_0\sigma^2\alpha^2 L}{2\theta_0(1-\rho(A_Q)^2)}\left[-d\log(1-\theta_0)+2\log(1/\zeta)\right],}$$

for $\theta_0 = \frac{\log(1/\zeta)}{d}\left[\sqrt{1+\frac{2d}{\log(1/\zeta)}}-1\right] < 1$, $\bar{u} = \min_{i\in\{1,..,d\}}\{u_i\}$ and an explicit M_0 under some generic assumptions †.

† The highlighted inequality holds for $c_i^2 + 4d_i \neq 0$, and more generic inequality holds for general choice of parameters.

Convergence rate results for TMM

- Let $\kappa = \dfrac{L}{\mu}$ and define the following sets[†]

$$\mathcal{S}_0 = \{(\vartheta,\psi) \mid \vartheta = 1 = \psi\}, \quad \mathcal{S}_+ = \left\{(\vartheta,\psi) \mid \psi > 1 \,\&\, 1 < \vartheta \le 2 - \frac{1}{\psi}\right\}, \; \mathcal{S}_- = \left\{(\vartheta,\psi) \mid 0 \le \psi < 1 \,\&\, \max\left\{2 - \frac{1}{\psi}, \frac{1}{1+\kappa(1-\psi)}\right\} \le \vartheta < 1\right\}$$

$$\mathcal{S}_1 = \left\{(\vartheta,\psi) \middle| \psi \ne 1, \left[1 - \sqrt{\frac{(1-\vartheta)\vartheta}{\kappa(1-\psi)}}\right]\left[1 - \frac{(1-\vartheta)(\mu\psi^2 - L(1-\psi)^2)}{L(1-\psi)\vartheta}\right] \le \left(1 - \frac{(1-\vartheta)\psi}{\kappa(1-\psi)}\right)^2\right\}.$$

- Consider TMM with parameters:

$$\beta_{\vartheta,\psi} = \frac{1 - \sqrt{\vartheta\alpha_{\vartheta,\psi}\mu}}{1 - \alpha_{\vartheta,\psi}\psi\mu}\left[1 - \sqrt{\frac{\alpha_{\vartheta,\psi}\mu}{\vartheta}}\right] \text{ and } \gamma_{\vartheta,\psi} = \psi\beta_{\vartheta,\psi} \;\text{ for } \alpha_{\vartheta,\psi} \in \begin{cases} \{\frac{1-\vartheta}{L(1-\psi)}\}, & \text{if } (\vartheta,\psi) \in \mathcal{S}_c := (\mathcal{S}_- \cup \mathcal{S}_+) \cap \mathcal{S}_1 \\ (0, \frac{1}{L}], & \text{if } (\vartheta,\psi) \in \mathcal{S}_0 \end{cases}, \qquad (1)$$

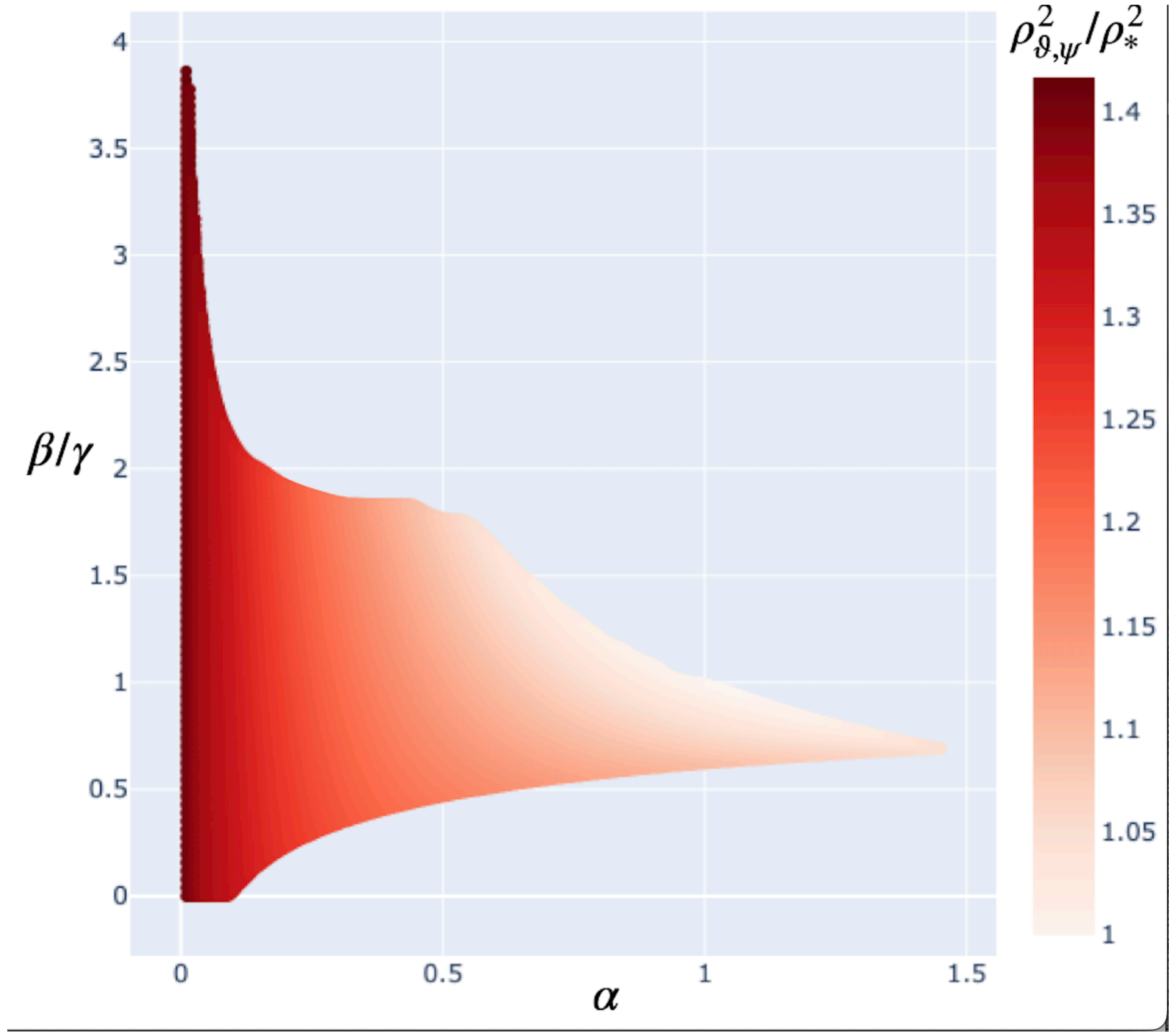
- **Theorem:** TMM without noise, with parameters $(\alpha_{\vartheta,\psi}, \beta_{\vartheta,\psi}, \gamma_{\vartheta,\psi}) \in \mathcal{S}_c \cup \mathcal{S}_0$ converges linearly at a rate

$$\rho^2_{\vartheta,\psi} = 1 - \sqrt{\vartheta\alpha_{\vartheta,\psi}\mu}.$$

[†] With the convention that $\max\{2 - \frac{1}{0}, \frac{1}{1+\kappa}\} = \frac{1}{1+\kappa}$

Reparametrizing TMM parameters

- The **FIRST** reparametrization of TMM with respect to two free variables.
- **Right figure:** The region $\alpha = \alpha_{\vartheta,\psi}$, $\psi = \gamma/\beta$ for $(\vartheta,\psi) \in \mathcal{S}_c$ where $L = 1$, $\mu = 0.1$, $x \in \mathbb{R}^d$, and the noise on the gradient is additive $\mathcal{N}(0, I_{10})$ and the comparison of rate $\rho^2_{\vartheta,\psi}$ with accelerated convergence rate $\rho^2_* = 1 - \sqrt{1/\kappa}$.



- $[\psi = 1]$, recovers **AGD:**

$$\beta = \gamma = \frac{1 - \sqrt{\alpha\mu}}{1 + \sqrt{\alpha\mu}} \text{ for } \alpha \in (0, \frac{1}{L}].$$

- $[\psi = 0]$ recovers **HB**:

$$\alpha = \frac{1 - \vartheta}{L}, \beta = \left[1 - \sqrt{\frac{\vartheta(1-\vartheta)}{\kappa}}\right]\left[1 - \sqrt{\frac{(1-\vartheta)}{\kappa\vartheta}}\right],$$

and $\gamma = 0$ for $\vartheta \in [\frac{\kappa}{\kappa+1}, 1)$.

Expected suboptimality of TMM on $f \in \mathcal{S}_\mu^L(\mathbb{R}^d)$

Theorem 4

The TMM on the objective $f \in \mathcal{S}_\mu^L(\mathbb{R}^d)$ where the gradient noise admits **Assumption 2** and the parameters are chosen as given in (1) satisfies

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \mathcal{O}(\rho_{\vartheta,\psi}^{2k}) + \left(\frac{\alpha_{\vartheta,\psi}(L\alpha_{\vartheta,\psi} + \vartheta)}{2(1-\rho_{\vartheta,\psi}^2)}\right) d\sigma^2,$$

where $\rho_{\vartheta,\psi}^2 = 1 - \sqrt{\vartheta\alpha_{\vartheta,\psi}\mu} < 1.$

- Theorem 5 implies the following convergence rates for other first order methods:

✦ **AGD**: $\rho_{\vartheta,\psi}^2 = \rho_\alpha^2 = 1 - \sqrt{\alpha\mu}$ where $\beta = \dfrac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ for $\alpha \in (0,1/L]$,

✦ **HB:** $\rho_{\vartheta,0}^2 = 1 - \sqrt{\dfrac{\vartheta(1-\vartheta)}{\kappa}}$ where $\alpha = \dfrac{1-\vartheta}{L}$, $\beta = \left[1 - \sqrt{\dfrac{\vartheta(1-\vartheta)}{\kappa}}\right]\left[1 - \sqrt{\dfrac{(1-\vartheta)}{\kappa\vartheta}}\right]$, and $\gamma = 0$ for $\vartheta \in [\dfrac{\kappa}{\kappa+1}, 1)$.

Entropic risk of TMM on $f \in \mathcal{S}_\mu^L(\mathbb{R}^d)$

Proposition 5

For $f \in \mathcal{S}_\mu^L(\mathbb{R}^d)$, assume the noise obeys **Assumption 2**. Then for $\theta < \theta_u^g$, we have

$$r_{k,\sigma^2}(\theta) < \frac{\sigma^2 d\alpha_{\vartheta,\psi}\big(\vartheta + \alpha_{\vartheta,\psi}L\big)}{(1-\bar{\rho}_{\vartheta,\psi}^2)(2-\theta\alpha_{\vartheta,\psi}\big(\vartheta+\alpha_{\vartheta,\psi}L\big))} + \mathcal{O}(\bar{\rho}_{\vartheta,\psi}^{2k}),$$

where $(\alpha_{\vartheta,\psi}, \beta_{\vartheta,\psi}, \gamma_{\vartheta,\psi})$ is chosen according to (1) and $\bar{\rho}_{\vartheta,\psi} \in (0,1)^\dagger$. Consequently,

$$r_{\sigma^2}(\theta) \leq \frac{\sigma^2 d\alpha_{\vartheta,\psi}\big(\vartheta + \alpha_{\vartheta,\psi}L\big)}{(1-\bar{\rho}_{\vartheta,\psi}^2)(2-\theta\alpha_{\vartheta,\psi}\big(\vartheta+\alpha_{\vartheta,\psi}L\big))}.$$

† We provide the explicit definitions of θ_u^g and $\bar{\rho}_{\vartheta,\psi}$ in the paper, and $\mathcal{O}(\,.\,)$ hides the terms that depends on initialization

EV@R of TMM on $f \in \mathcal{S}_{\mu}^{L}(\mathbb{R}^d)$

Theorem 6 (Informal)

Consider the noisy TMM to minimize the objective $f \in \mathcal{S}_{\mu}^{L}(\mathbb{R}^d)$ under the setting of Proposition 5. Let $\varphi \in (0,1)$ be fixed. Set $\theta_\varphi = \varphi \theta_u^g$ and define

$$\bar{E}_{1-\zeta}(\vartheta, \psi) = \begin{cases} \dfrac{\sigma^2 \alpha_{\vartheta,\psi}(\vartheta + \alpha_{\vartheta,\psi} L)}{2} \left(\sqrt{\dfrac{d}{1 - \bar{\bar{\rho}}_{\vartheta,\psi}}} + \sqrt{2 \log(1/\zeta)} \right)^2, & \text{if } \zeta < \zeta_0, \\ \dfrac{\sigma^2 d \alpha_{\vartheta,\psi}(\vartheta + \alpha_{\vartheta,\psi} L)}{(1 - \bar{\bar{\rho}}_{\vartheta,\psi})(2 - \theta_\varphi^g \alpha_{\vartheta,\psi}(\vartheta + \alpha_{\vartheta,\psi} L))} + \dfrac{2\sigma^2}{\theta_\varphi^g} \log(1/\zeta), & \text{otherwise}, \end{cases}$$

for some $\bar{\bar{\rho}}_{\vartheta,\psi} \in (0,1)$ and ζ_0 we explicitly provide, then EV@R admits the bound

$$EV@R_{1-\zeta}[f(x_k) - f(x_*)] \leq \bar{E}_{1-\zeta}(\vartheta, \psi) + \mathcal{O}((\bar{\bar{\rho}}_{\vartheta,\psi})^k)$$

† We provide the explicit definitions of $\bar{\bar{\rho}}_{\vartheta,\psi}$ in the paper, and $\mathcal{O}(.)$ hides the terms that depends on initialization.

Tail bounds for TMM on $f \in \mathcal{S}_{\mu}^{L}(\mathbb{R}^d)$

- Theorem 6 implies

$$\mathbb{P}\left\{f(x_k) - f(x_*) \geq t_\zeta\right\} < \exp\left\{\frac{\theta}{2\sigma^2}\bar{\rho}_{\vartheta,\psi}^{2k}\mathscr{V}_0 - t_\zeta + \frac{\theta d\alpha_{\vartheta,\psi}(\vartheta + \alpha_{\vartheta,\psi}L)}{2(1-\bar{\rho}_{\vartheta,\psi}^2)(2-\theta\alpha_{\vartheta,\psi}(\vartheta+\alpha_{\vartheta,\psi}L))}\right\},$$

- where $\mathscr{V}_0$ depends on initialization†.

† We give the explicit form of $\mathscr{V}_0$ in the paper.

Experiments: Risk-averse TMM on quadratic objectives

- Consider the quadratic objective:

$$f(x) = \frac{1}{2}x^\top Qx + b^\top x + 2.5\|x\|^2,$$

where $b = \frac{1}{\|\tilde{b}\|^2}\tilde{b}$ for $\tilde{b} = [1,....,1] \in \mathbb{R}^{10}$, $Q = Diag_{i=1,..,10}(i^2)$, and variance of the noise is $\sigma^2 = 1$.

- Parameters $(\alpha_q, \beta_q, \gamma_q)$ of **risk-averse TMM (RA-TMM)**: Solve

$$(\alpha_q, \beta_q, \gamma_q) = \underset{(\alpha,\beta,\gamma)\in\mathcal{S}_q}{\operatorname{argmin}} \bar{E}^q_{1-\zeta}(\alpha, \beta, \gamma)$$

$$\text{s.t.} \quad \frac{\rho^2(\alpha, \beta, \gamma)}{\rho^2_{q,*}} \leq (1+\epsilon),$$

using grid-search, where $\rho_{q,*} = 1 - \frac{2}{\sqrt{3\kappa+1}}$, $\zeta = 0.95$ confidence level, and $\epsilon = 0.25$.

- For risk-averse AGD (RA-AGD), we added the constraint $\beta = \gamma$ to the problem above.

Rutgers Business School
Newark and New Brunswick

Experiments: Risk-averse TMM on quadratic objectives

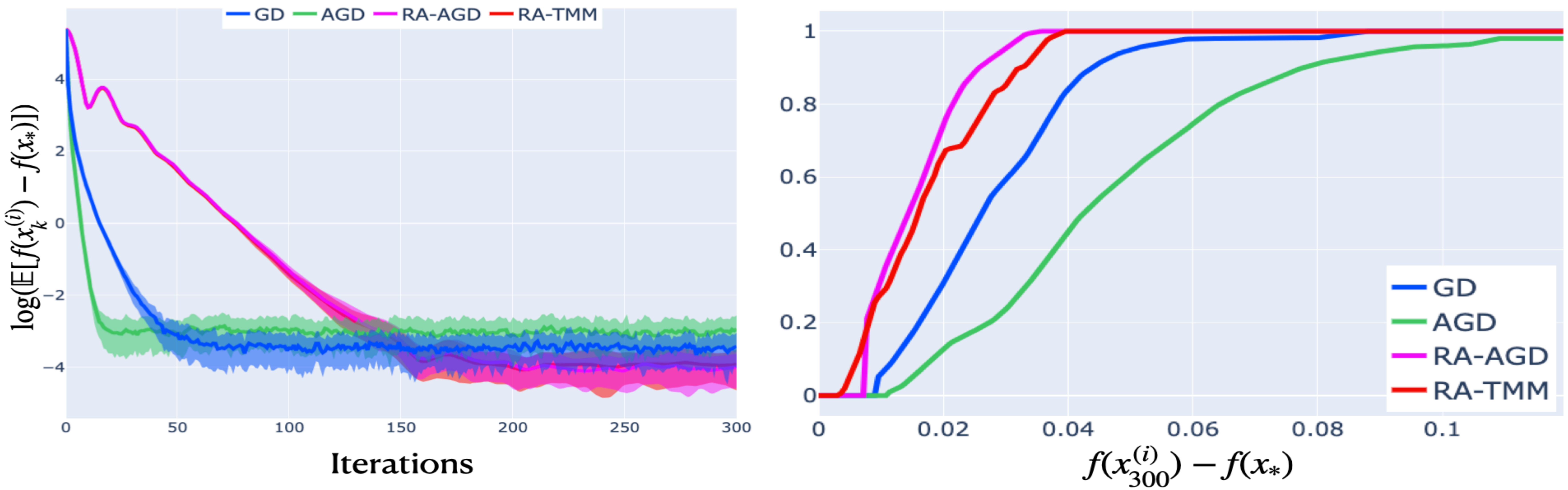


Figure: (Left) The expected suboptimality versus iterations for GD, AGD, RA-AGD and RA-TMM. **(Right)** The cumulative distribution of the suboptimality of the last iterates for GD, AGD, RA-AGD and RA-TMM after $k = 300$ iterations on the quadratic loss function.

- We plot the average $(\bar{f}_1, \ldots, \bar{f}_{300})$ where $\bar{f}_k := \frac{1}{50}\sum_{i=1}^{50} f(x_k^{(i)}) - f(x_*)$ over the samples $\{x_k^{(i)}\}_{i=1}^{50}$.
- We highlight the region between $(\bar{f}_0 \pm \sigma_0^f, \ldots, \bar{f}_{300} \pm \sigma_{300}^f)$ where $\sigma_k^f := \left(\frac{1}{50}\sum_{i=1}^{50} |f(x_k^{(i)}) - f(x_*)|^2\right)^{1/2}$.

Experiments: Risk-averse TMM on quadratic objectives

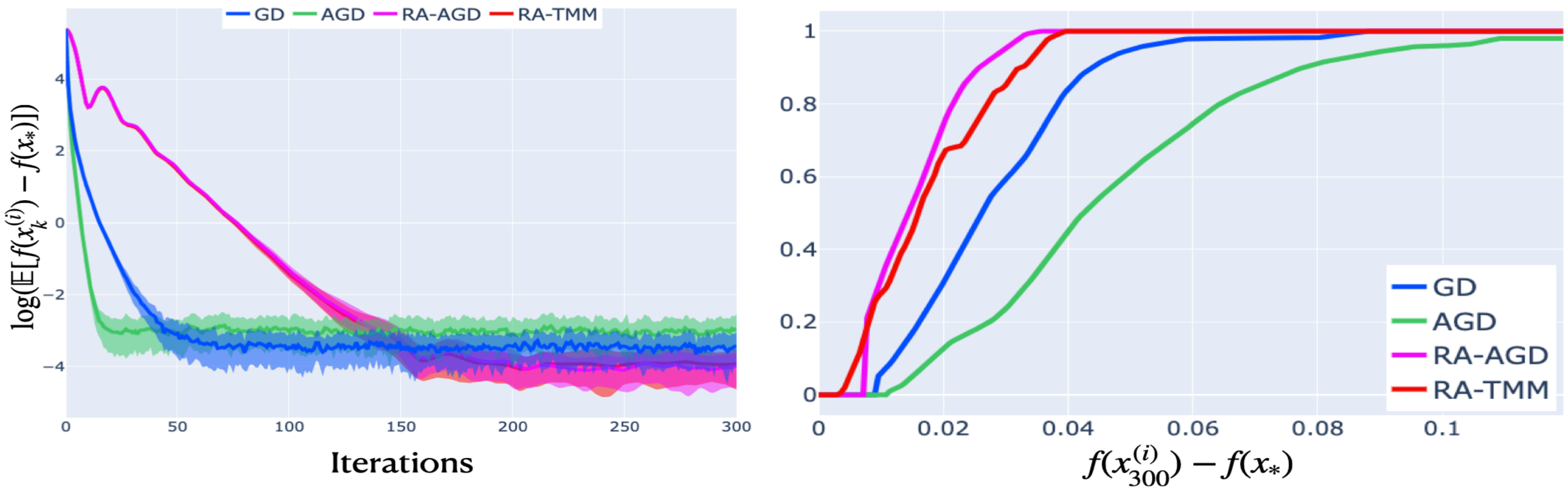


Figure: (Left) The expected suboptimality versus iterations for GD, AGD, RA-AGD and RA-TMM. **(Right)** The cumulative distribution of the suboptimality of the last iterates for GD, AGD, RA-AGD and RA-TMM after $k = 300$ iterations on quadratic loss function.

- Risk-averse algorithms trades convergence rate with entropic risk.
- The distribution of $\{f(x_{300}) - f(x_*)\}$ of risk-averse algorithms stochastically dominates the one of standard algorithms

Experiments: Risk-averse TMM on logistic regression

- We design risk-averse TMM (RA-TMM) for logistic loss:

$$f(x) = \sum_{i=1}^{N} \frac{1}{N} f_i(x) := \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp\{-y_i(X_i^\top x)\}) + \frac{1}{2}\|x\|^2,$$

where $X_i \in \mathbb{R}^{100}$ is the feature vector and $y_i \in \{-1,1\}$ is the label of i-th sample, with $d = 100$, $N = 1000^\dagger$.

- Parameters $(\alpha_{\vartheta_*,\psi_*}, \beta_{\vartheta_*,\psi_*}, \gamma_{\vartheta_*,\psi_*})$ of **RA-TMM:** Solve

$$(\vartheta_*, \psi_*) := \underset{(\vartheta,\psi)\in\mathcal{S}_c\cup\mathcal{S}_0}{\operatorname{argmin}} \ \bar{E}_{1-\zeta}(\vartheta, \psi)$$

$$\text{s.t.} \quad \frac{\rho^2_{\vartheta,\psi}}{\rho^2_*} \leq (1 + \epsilon),$$

using grid-search, where $\rho^2_* = 1 - \sqrt{1/\kappa}$, $\zeta = 0.95$, and $\epsilon = 0.25$.

- For risk-averse AGD (RA-AGD), we added the constraint $\beta = \gamma$ to the problem above.

† See the paper for further details on the generation of synthetic data X and y.

Experiments: Risk-averse TMM on logistic regression

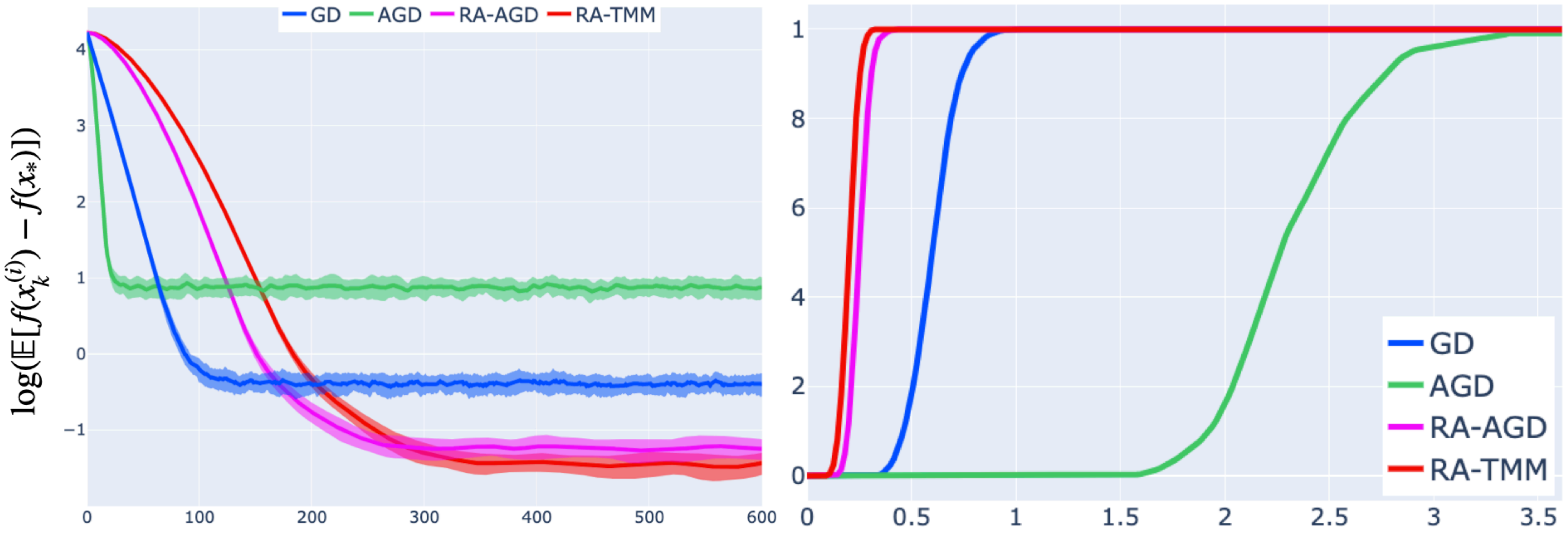


Figure: (Left) The expected suboptimality versus iterations for GD, AGD, RA-AGD and RA-TMM. **(Right)** The cumulative distribution of the suboptimality of the last iterates for GD, AGD, RA-AGD and RA-TMM after $k = 600$ iterations on logistic regression where the noise is $\mathcal{N}(0, I_{100})$.

- We plot the average $(\bar{f}_1, \ldots, \bar{f}_{300})$ where $\bar{f}_k := \frac{1}{50}\sum_{i=1}^{50} f(x_k^{(i)}) - f(x_*)$ over the samples $\{x_k^{(i)}\}_{i=1}^{50}$.
- We highlight the region between $(\bar{f}_0 \pm \sigma_0^f, \ldots, \bar{f}_{600} \pm \sigma_{600}^f)$ where $\sigma_k^f := \left(\frac{1}{50}\sum_{i=1}^{50} |f(x_k^{(i)}) - f(x_*)|^2\right)^{1/2}$.

Experiments: Risk-averse TMM on logistic regression

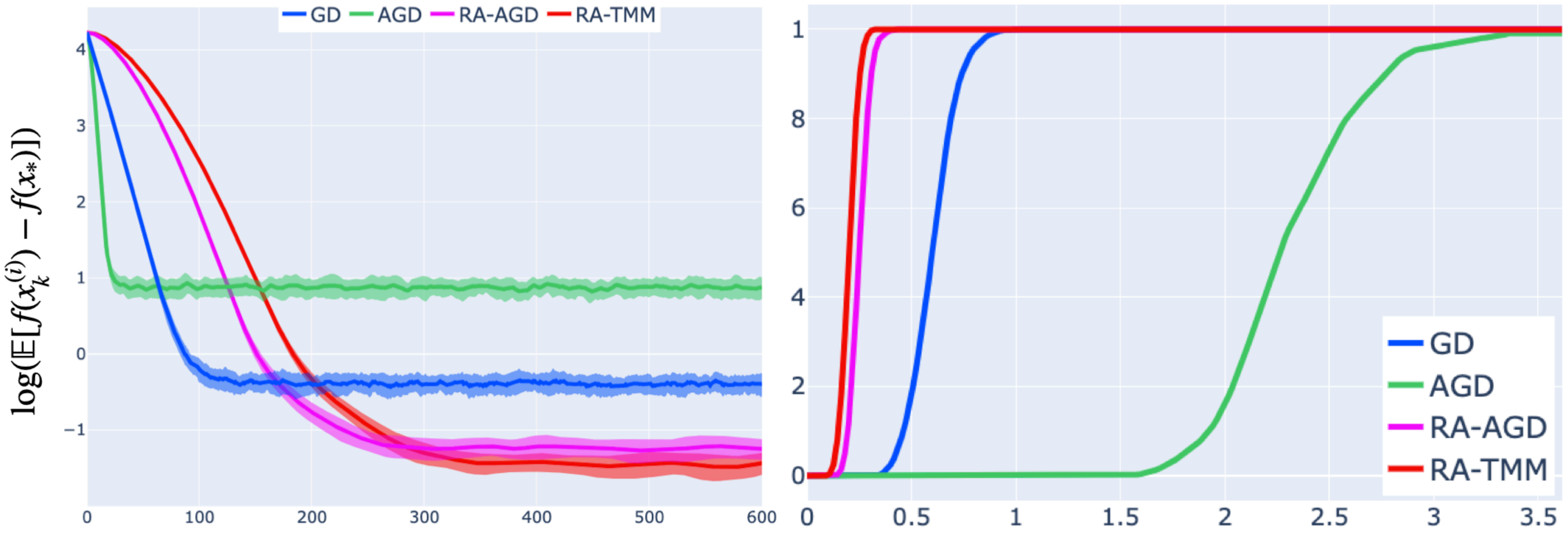


Figure: (Left) The expected suboptimality versus iterations for GD, AGD, RA-AGD and RA-TMM. **(Right)** The cumulative distribution of the suboptimality of the last iterates for GD, AGD, RA-AGD and RA-TMM after $k = 600$ iterations on logistic regression where the noise is $\mathcal{N}(0, I_{100})$.

- Our risk-averse TMM algorithms trade convergence rate with entropic risk.
- The distribution of $\{f(x_{600}) - f(x_*)\}$ of risk-averse algorithms stochastically dominates that of GD/AGD with standard parameters.

Saddle point problems

Empirical risk minimization (ERM)

$$\min_{x\in\mathbb{R}^d} \mathbb{E}[f(x)] = \min_{x\in\mathbb{R}^d} \sum_{i=1}^{N} \frac{1}{N} f_i(x),$$

where N is the sample size.

Distributionally robust ERM

$$\min_{x\in\mathbb{R}^d} \sup_{Q\in\mathscr{P}} \mathbb{E}_Q[f(x)] = \min_{x\in\mathbb{R}^d} \max_{y\in\mathscr{P}_{r,n}} \mathscr{L}(x,y) := \sum_{i=1}^{N} y_i f_i(x),$$

where $\mathscr{P}$ is an uncertainty set around empirical dist. and

$$\mathscr{P}_{r,n} := \{y \in \mathbb{R}^N : y^\top 1 = 1,\, y \geq 0,\, D_{KL}(y\|1/N) \leq r/N\}.$$

Saddle point problems

Empirical risk minimization (ERM)

$$\min_{x\in\mathbb{R}^d} \mathbb{E}[f(x)] = \min_{x\in\mathbb{R}^d} \sum_{i=1}^{N} \frac{1}{N} f_i(x),$$

where N is the sample size.

Distributionally robust ERM

$$\min_{x\in\mathbb{R}^d} \sup_{Q\in\mathscr{P}} \mathbb{E}_Q[f(x)] = \min_{x\in\mathbb{R}^d} \max_{y\in\mathscr{P}_{r,n}} \mathscr{L}(x,y) := \sum_{i=1}^{N} y_i f_i(x),$$

where $\mathscr{P}$ is an uncertainty set around empirical dist. and

$\mathscr{P}_{\rho,n} := \{y \in \mathbb{R}^N : y^\top 1 = 1,\, y \geq 0,\, D_{KL}(y\|1/N) \leq \rho/N\}$.

- Strongly convex strongly concave (SCSC) saddle point (SP) problem:

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \mathscr{L}(x,y),$$

where $\mathscr{L}$ is smooth and strongly convex in x and strongly concave in y.

- SCSC SP arise in
  - ✦ Robust training of ML models,
  - ✦ Robust optimization,
  - ✦ Designing fair classifiers [Nouiehed, 2019],
  - ✦ Constrained optimization (via Lagrangian duality).

Stochastic accelerated primal and dual algorithm

- The stochastic accelerated primal dual algorithm (SAPD)† [Zhang, Aybat, Gurbuzbalaban, 2021]

$$\tilde{q}_k = (1+\theta)\,\tilde{\nabla}_y \mathcal{L}(x_k, y_k) - \theta\,\tilde{\nabla}_y \mathcal{L}(x_{k-1}, y_{k-1}),$$

$$y_{k+1} = y_k + \delta \tilde{q}_k,$$

$$x_{k+1} = x_k - \tau\,\tilde{\nabla}_x \mathcal{L}(x_k, y_{k+1}),$$

- Pareto-optimal parameter design trading rate with robustness [Zhang, Aybat, Gurbuzbalaban, 2021]

Assumption 3: Let $\{x_k, y_k\}$ be SAPD iterates, then gradient estimates satisfy

- $\mathbb{E}[\,\tilde{\nabla}_y f(x_k, y_k) - \nabla_y f(x_k, y_k) \mid x_k, y_k] = 0$ and $\mathbb{E}[\,\tilde{\nabla}_x f(x_k, y_k) - \nabla_x f(x_k, y_k) \mid x_k, y_k] = 0$,
- $\tilde{\nabla}_y f$ and $\tilde{\nabla}_x f$ are **independent** from each other,
- $\exists \sigma_{(p)} > 0$ s.t. $\mathbb{E}[\|\tilde{\nabla}_y f(x_k, y_k) - \nabla_y f(x_k, y_k)\|^p \mid x_k, y_k] \le \sigma_{(p)}^p$ & $\mathbb{E}[\|\tilde{\nabla}_x f(x_k, y_k) - \nabla_x f(x_k, y_k)\|^p \mid x_k, y_k] \le \sigma_{(p)}^p$, $p \in \{2,3,4\}$
- $\tilde{\nabla}_x f - \nabla_x f$ and $\tilde{\nabla}_y f - \nabla_y f$ are **stationary,** and **independent from the past**.

†With a slight abuse of notation, we use θ as the algorithm parameter to be consistent with [Zhang & Aybat. 2019]

Variance-Reduced SAPD (VR-SAPD)

- Let $\mathscr{L}$ be SCSC function of SP problem with solution (x_*, y_*).
- Introduce $\xi_k^{(\theta)} = [(x_k^{(\theta)})^\top, (y_k^{(\theta)})^\top, (x_{k-1}^{(\theta)})^\top, (y_{k-1}^{(\theta)})^\top]^\top$, where $[(x_k^{(\theta)})^\top, (y_k^{(\theta)})^\top]^\top$ are generated by SAPD with parameters:

$$\tau_\theta = \frac{1-\theta}{\mu_x}, \delta_\theta = \frac{1-\theta}{\mu_y \theta}, \theta \in [\hat{\theta}, 1) \text{ for some } \hat{\theta} \in (0,1)^\dagger.$$

Theorem (informal) [Can, Aybat, Gurbuzbalaban, 2022]

Under **Assumption 3** when variance $\sigma_{(2)}^2$ is "small enough", the stationary distribution exists†† and we have

$$\lim_{k\to\infty} \mathbb{E}[\xi_k^{(\theta)}] = \xi_* + (1-\theta)(\nabla^{(2)}\mathscr{L}_*)^{-1}\left(\nabla^{(3)}\mathscr{L}_* M_w\right) + \mathcal{O}((1-\theta)^{3/2}), \tag{2}$$

where $\nabla^{(2)}\mathscr{L}_*$ is the Hessian, $\nabla^{(3)}\mathscr{L}_*$ third-order tensor appears in Taylor expansion around (x_*, y_*), and M_w is a fixed matrix that we can characterize.

- Using Richardson-Romberg extrapolation and characterization of stationary distribution mean (2), we introduce the variance-reduced SAPD.

† The function $\mathscr{L}$ is μ_x strongly convex and μ_y strongly convcave
†† Using the techniques provided in [Hairer, 2008]

Rutgers Business School
Newark and New Brunswick

Our contributions [Can, Aybat, Gurbuzbalaban, 2022]

- Let $\bar{\xi}_k^{(\theta)} = \frac{1}{k}\sum_{i=1}^{k} \xi_k^{(\theta)}$, the VR-SAPD calculates the sequence $\{2\bar{\xi}_k^{\theta} - \bar{\xi}_k^{(2\theta-1)}\}$ for θ, $(2\theta-1) \in [\hat{\theta},1)$:

VR - SAPD	SAPD
$\lim_{k\to\infty} \mathbb{E}[2\bar{\xi}_k^{(\theta)} - \bar{\xi}_k^{(2\theta-1)}] = \xi_* + \mathcal{O}((1-\theta)^{3/2})$	$\lim_{k\to\infty} \mathbb{E}[\xi_k^{(\theta)}] = \xi_* + \bar{\mathcal{O}}((1-\theta)) + \mathcal{O}((1-\theta)^{3/2})$

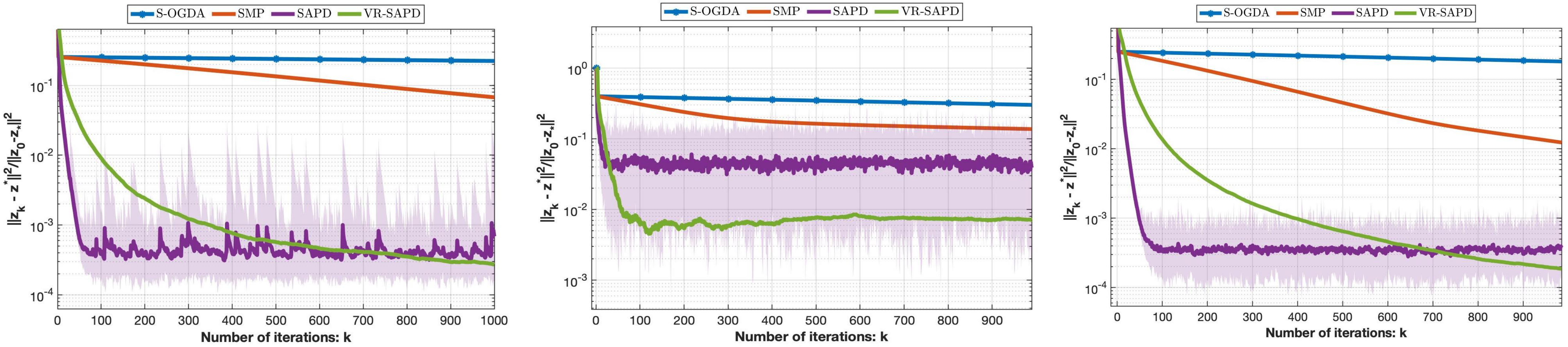


Figure: Comparison of S-OGDA, SMP, SAPD, and VR-SAPD on MNIST, DryBean, and Arcene (from left to right) on empirical DRO problem in terms of the relative expected distance squared $\mathbb{E}[\|z_k - z_*\|^2]/\|z_0 - z_*\|^2$.

Summary

``More risk, more (expected) reward" , Folklore

- There are fundamental trade-offs (rate vs robustness to noise/risk) when designing a first-order algorithm.
  - Heisenberg-like impossibility results.
  - First-time rate/risk results for Triple Momentum Methods (improved heavy-ball analysis)
  - Similar trade-offs for min-max optimization.
- Introduced **"Risk Averse Momentum Methods"**
  - On the Pareto-optimal curve trading rate with risk/robustness to noise.
  - Results in better tail behavior for suboptimality.
- We obtain stronger guarantees (conv. rate to the stationary distribution)
  - Wasserstein distances for translating deterministic convergence analysis to the stochastic case.
  - This can be used to debias the stationary distribution/improve the performance.

Thank you

References

- [Entropic Risk-Averse Generalized Momentum Methods, 2022] (with B. Can), Submitted.
- [Variance Reduced Saddle Point Algorithms, 2022] (with B. Can and N.S. Aybat), Submitted.
- [Accelerated Linear Convergence of Stochastic Momentum Methods in Wasserstein Distances, 2019] (with B. Can and L. Zhu), International Conference on Machine Learning.
- [Decentralized Langevin dynamics and Hamiltonian Monte Carlo] (with Y. Hu, X. Gao and L. Zhu), JMLR 2021.
- [Robust Accelerated Gradient Methods for Strongly Convex Functions, 2019] (with N.S. Aybat, A. Fallah, A. Ozdaglar), Siam Journal on Optimization.

Acknowledgements

Special thanks to all my collaborators, and funding from NSF CCF 1814888, ONR N00014-21-1-2244.