Clipped Stochastic Methods for Variational Inequalities with Heavy-Tailed Noise

Eduard Gorbunov*[1,2,3], Marina Danilova*[1], David Dobre*[2], Pavel Dvurechensky[4], Alexander Gasnikov[1,5,6], Gauthier Gidel[2,7]

[1]MIPT, [2]Mila & UdeM, [3]MBZUAI, [4]WIAS, [5]HSE University, [6]IITP RAS, [7]Canada CIFAR AI Chair

1. Preliminaries

Problem: stochastic unconstrained variational inequality problem (VIP)

$$\text{find } x^* \in \mathbb{R}^d \quad \text{such that} \quad F(x^*) = 0$$

$$F(x) = \mathbb{E}\left[F_\xi(x)\right]$$

- Information about the problem is available through the stochastic oracle calls $F_\xi(x)$
- Examples include stochastic/finite-sum min-max and minimization problems

Assumptions: all conditions are required only on some ball around the solution, i.e., for all $x, y \in B_r(x^*)$, where $B_r(x^*) = \{x \in \mathbb{R}^d \mid \|x - x^*\| \leq r\}$ and $r \sim R_0 = \|x^0 - x^*\|$

- Bounded variance / heavy (non-sub-Gaussian) tails

$$\mathbb{E}\left[\|F_\xi(x) - F(x)\|^2\right] \leq \sigma^2$$

We consider 6 different classes of problems (4 of them allow non-monotone problems). Each class is defined by 1-2 of the conditions below.

- Lipschitzness **(Lip)** $\qquad \|F(x) - F(y)\| \leq L\|x - y\|$
- Monotonicity **(Mon)** $\qquad \langle F(x) - F(y), x - y\rangle \geq 0$
- Star-Negative Comonotonicity **(SNC)**

$$\langle F(x), x - x^*\rangle \geq -\rho\|F(x)\|^2, \quad \rho \in [0, +\infty)$$

When $\rho = 0$ the operator is called Star-Monotone **(SM)**

- Quasi-Strong Monotonicity **(QSM)**

$$\langle F(x), x - x^*\rangle \geq \mu\|x - x^*\|^2, \quad \mu \geq 0$$

- Star-Cocoercivity **(SC)** $\qquad \|F(x)\|^2 \leq \ell\langle F(x), x - x^*\rangle$

$$\begin{array}{ccccc} (\text{Mon}) & \Longrightarrow & (\text{SM}) & \Longrightarrow & (\text{SNC}) \\ & & \Uparrow & \nearrow & \Uparrow \\ & & (\text{QSM}) & \xRightarrow{+(\text{Lip})} & (\text{SC}) \end{array}$$

Relation between the assumptions on the structured non-monotonicity of the problem

High-probability guarantees: $\qquad \mathbb{P}\{\text{Metric} \leq \varepsilon\} \geq 1 - \beta$

- Possible metrics: $Gap_R(x) = \max_{y \in B_R(x^*)}\langle F(y), x - y\rangle$, $\|F(x)\|^2$, $\|x - x^*\|^2$
- Sensitive to the noise distrib. → more accurately describe the methods' behavior

2. Our Contributions

- New high-probability results for variational inequalities with heavy-tailed noise
  - ✓ Allow heavy-tailed noise
  - ✓ Unconstrained problem
- Tight analysis
  - ✓ Logarithmic dependence on $1/\beta$
  - ✓ Optimal (up to logarithms) results
  - ✓ Improvement upon previous best-known result under the light-tails assumption
- Weak assumptions
  - ✓ All assumptions are made just on a ball around the solution
  - ✓ Key ingredient: we prove that the considered algorithms do not leave this ball with high-probability
- Numerical experiments
  - ✓ Empirically observed heavy-tailed noise in GANs training
  - ✓ Showed that gradient clipping significantly improves the results

3. Clipped SEG and SGDA

We consider standard Stochastic Extragradient (SEG) and Stochastic Gradient Descent-Ascent (SGDA) with clipping of the update vectors

Clipped Stochastic Extragradient (clipped-SEG)

extrapolation step: $$\widetilde{x}^k = x^k - \gamma_1 \cdot \text{clip}\left(F_{\xi_1^k}\left(x^k\right), \lambda_{1,k}\right)$$

update step: $$x^{k+1} = x^k - \gamma_2 \cdot \text{clip}\left(F_{\xi_2^k}\left(\widetilde{x}^k\right), \lambda_{2,k}\right)$$

Clipped Stochastic Gradient Descent-Ascent (clipped-SGDA)

update step: $$x^{k+1} = x^k - \gamma \cdot \text{clip}\left(F_{\xi^k}\left(x^k\right), \lambda_k\right)$$

- Clipping operator: $\text{clip}(x, \lambda) = \min\left\{1, \frac{\lambda}{\|x\|}\right\}x$, λ – clipping level
- Clipping levels are properly chosen: effect of heavy-tailed noise is reduced, while the bias is not too large
- clipped-SEG: $\xi_{1,k}, \xi_{2,k}$ are i.i.d. samples independent from prev. steps, $\gamma_2 \leq \gamma_1$
- clipped-SGDA: ξ_k is independent from previous steps

Summary of the Complexity Results

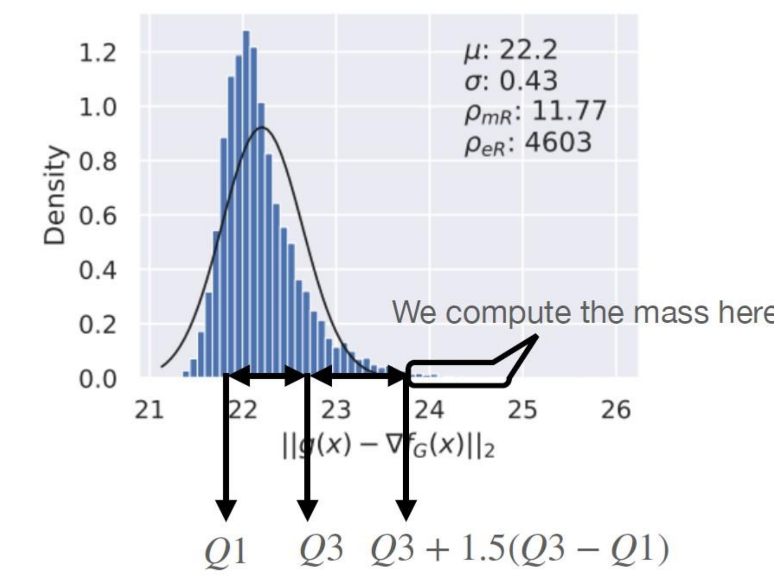
Setup	Method	Metric	Complexity	HT?	UD?
(Mon)+(Lip)	Mirror-Prox [Juditsky et al., 2011a]	$\texttt{Gap}_D(\widetilde{x}^K_{\text{avg}})$	$\max\left\{\frac{LD^2}{\varepsilon}, \frac{\sigma^2 D^2}{\varepsilon^2}\right\}$	✗	✗
	clipped-SEG	$\texttt{Gap}_R(\widetilde{x}^K_{\text{avg}})$	$\max\left\{\frac{LR^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	✓	✓
(SNC)+(Lip)	clipped-SEG	$\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2$	$L^2\max\left\{\frac{R^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	✓	✓
(QSM)+(Lip)	clipped-SEG	$\|x^K - x^*\|^2$	$\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	✓	✓
(Mon)+(SC)	clipped-SGDA	$\texttt{Gap}_R(x^K_{\text{avg}})$	$\max\left\{\frac{\ell R^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	✓	✓
(SC)	clipped-SGDA	$\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2$	$\ell^2\max\left\{\frac{R^2}{\varepsilon}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}$	✓	✓
(QSM)+(SC)	clipped-SGDA	$\|x^K - x^*\|^2$	$\max\left\{\frac{\ell}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\right\}$	✓	✓

- Logarithmic factors of $1/\varepsilon$ and $1/\beta$ are omitted
- HT? = Heavy-Tailed noise?
- UD? = Unbounded domain?

4. Numerical Experiments

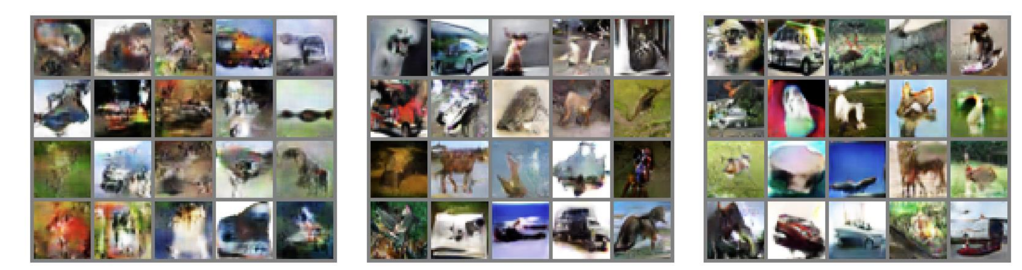
Training WGAN-GP on CIFAR10

1. WGAN-GP on CIFAR10 has heavy-tailed gradients

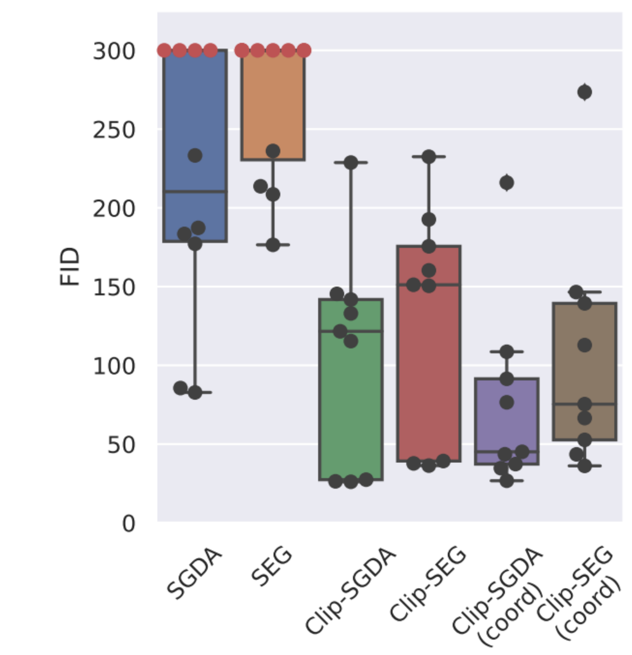


- ρ_{mR}: relative fraction of mass after $Q_3 + 1.5 \cdot (Q_3 - Q_1)$
  - For normal distr.: ≈ .35 % of the mass
  - In this plot: ≈ 12 times more
- ρ_{eR}: relative fraction of mass after $Q_3 + 3 \cdot (Q_3 - Q_1)$
  - For normal distr.: ≈ 10^{-4}% of the mass
  - In this plot: ≈ 4603 times more

2. Clipping helps for WGAN-GP on CIFAR10



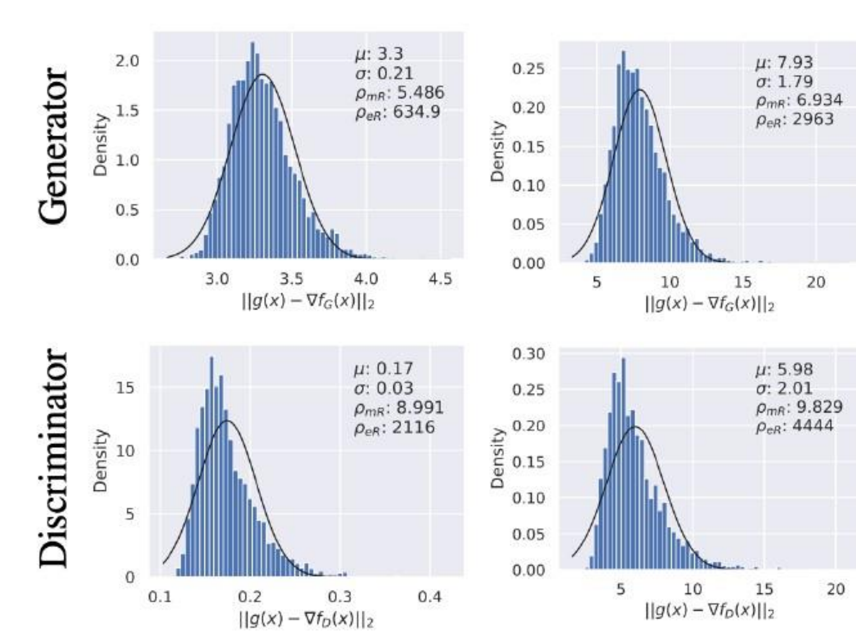
(a) SGDA (67.4) (b) clipped-SGDA (19.7) (c) clipped-SEG (25.3)



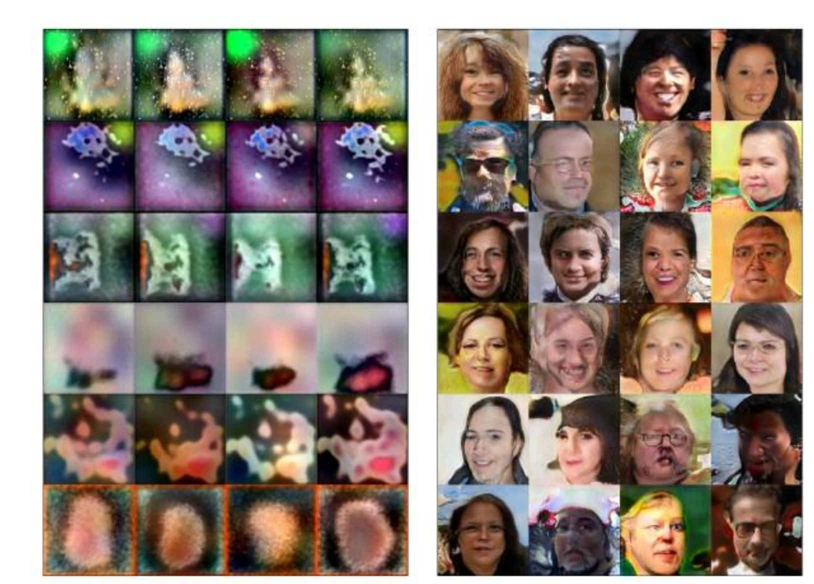
- Clipping improves the results both in terms of the quality of the generated images and Fréchet inception distance (FID)
- Methods without clipping diverge for most of the tested hyperparameters

Training StyleGAN2 on FFHQ

1. StyleGAN2 on FFHQ has heavy-tailed gradients
2. Clipping helps for StyleGAN2 on FFHQ



(a) Initialization (b) clipped-SGDA



(c) SGDA (d) clipped-SGDA

- Still not matching Adam (on this GAN)
- StyleGAN2 is full of tricks and heuristics
- It has been tuned for Adam

References

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. Stochastic Systems, 1(1):17–58, 2011a.