Corner Gradient Descent: Provable acceleration of power-law convergence of SGD[1]

Dmitry Yarotsky[2]

Peking University
FoMI: A Seminar on the Foundations of Machine Intelligence
March 26, 2026

[1] https://openreview.net/forum?id=nOXCfIdhD9 (ICLR 2026)

[2] relies on earlier work with Maksim Velikanov

1. Overview

2. In-depth exposition

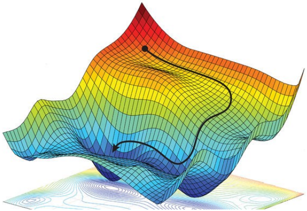
1. Overview

2. In-depth exposition

Gradient descent

Basic algorithm for learning (optimizing) neural networks and other predictive models

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla L(\mathbf{w}_t)$$



(https://reconsider.news/2018/05/09/ai-researchers-allege-machine-learning-alchemy/)

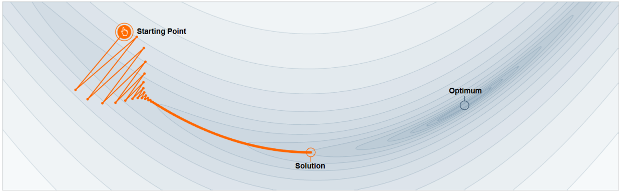
Stochastic Gradient Descent (SGD): instead of exact loss $L(\mathbf{w}_t)$ use its estimates $L_{B_t}(\mathbf{w}_t)$ computed over mini-batches B_t

Heavy Ball[3]

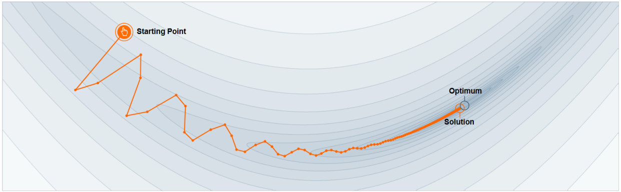
(S)GD can be accelerated by exploiting "momentum vector" $\mathbf{u}_t$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{u}_{t+1}, \quad \mathbf{u}_{t+1} = \alpha \nabla L(\mathbf{w}_t) + \beta \mathbf{u}_t$$

Without momentum



With momentum



(https://distill.pub/2017/momentum/)

[3] Polyak (1964)

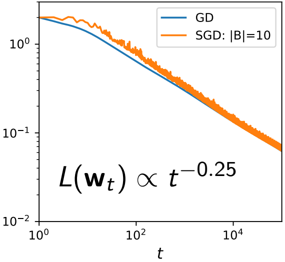
Power laws

Optimization of neural networks: a high-dimensional and ill-conditioned problem

Theoretically and practically, optimization trajectories in these problems are often well described by **power laws**:

$$L(\mathbf{w}_t) \propto t^{-\zeta}$$

For example, for MNIST hand-written digit classification $\zeta \approx 0.25$

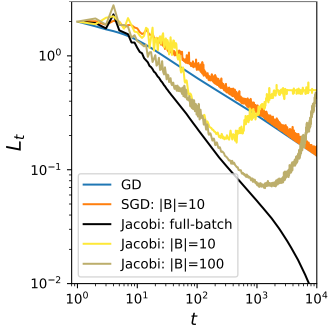


Acceleration of convergence exponents

For **non-stochastic** GD, momentum with a Jacobi schedule $\beta_t \sim 1 - \frac{const}{t}$ allows to **double** the convergence exponent[4]:

$$L(\mathbf{w}_t) = O(t^{-\mathbf{2}\zeta})$$

But in **Stochastic** GD such acceleration only works for a limited number of iterations; after that optimization **diverges**



The challenge: Can we achieve a stable acceleration for Stochastic GD?

A solution: Corner Gradient Descent

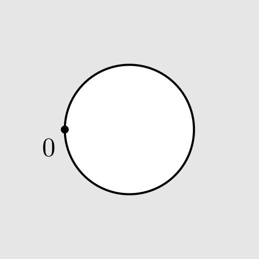
[4] Nemirovskiy & Polyak (1984), Brakhage (1987)

Generalized (S)GD's as contours in $\mathbb{C}$

Stationary generalizations of (S)GD with arbitrary linear memory can be identified with **contours** $\gamma \subset \mathbb{C}$ through **frequency response function** Ψ: $\gamma = \Psi(\{|z| = 1\})$

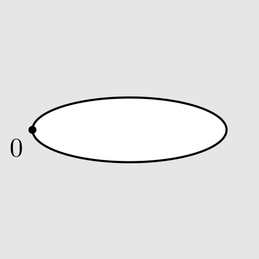
Plain (S)GD:

a circle



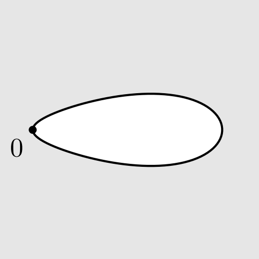
Heavy Ball:

an ellipse



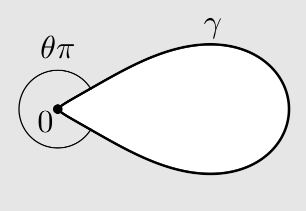
General memory-1:

a Zhukovsky airfoil



Corner algorithms

Correspond to contours with external angle $\theta\pi$



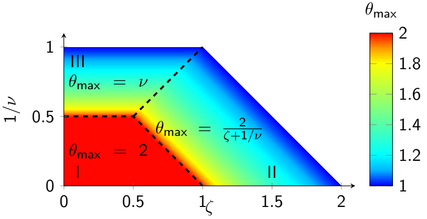
Accelerate convergence exponent of **non-stochastic** GD by the factor θ:

$$L(\mathbf{w}_t) \propto t^{-\zeta} \quad \rightsquigarrow \quad L(\mathbf{w}_t) \propto t^{-\theta\zeta}$$

Corner algorithms in mini-batch stochastic setting

But corner algorithms also **amplify sampling noise**

For tasks with power-law spectral data, maximum acceleration $\theta_{\max}$ is obtained by balancing deterministic acceleration and noise amplification



Finite-memory algorithms

Ideal corner algorithms require **infinite memory**, but can be efficiently approximated by finite-memory algorithms thanks to fast rational approximations of power functions[5]:

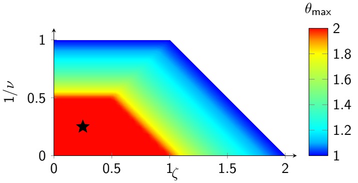
$$\sup_{\Re(z)>0, |z|<1} \left| z^\theta - \frac{P(z)}{Q(z)} \right| = O\left(e^{-c\sqrt{\deg(PQ)}} \right)$$

[5] Newman (1964), Gopal & Trefethen (2019)

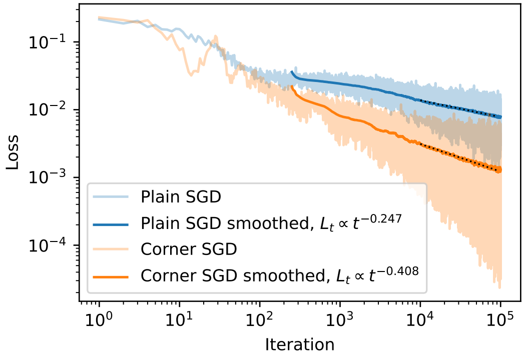
Example 1[6]: Indicator function $\mathbf{1}_{[1/4,3/4]}(x)$

ReLU NN with one hidden layer; only output layer is trained

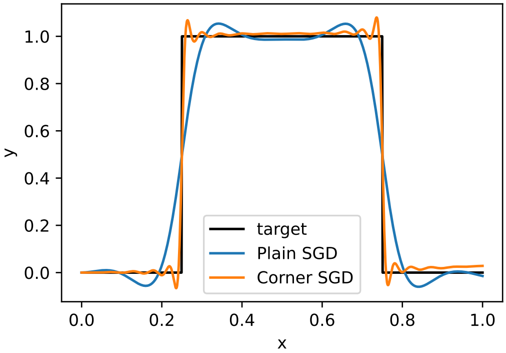
Theory: $\zeta = \frac{1}{4}$; feasible accelerations up to $\theta_{\max} = 2$



Loss during training



Target and predictions

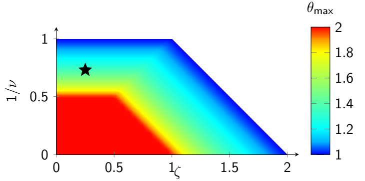


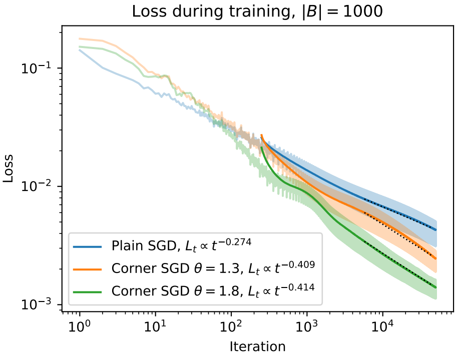
[6] Demo: https://github.com/yarotsky/corner-gradient-descent

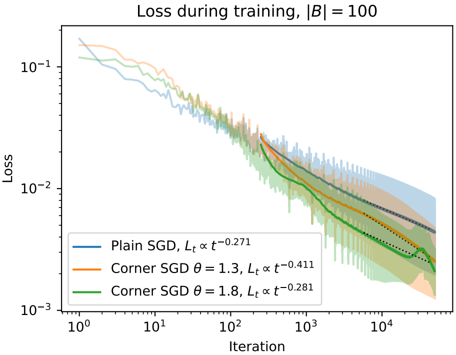
Example 2: classifier MNIST

NN with one hidden layer, $\zeta \approx 0.25$

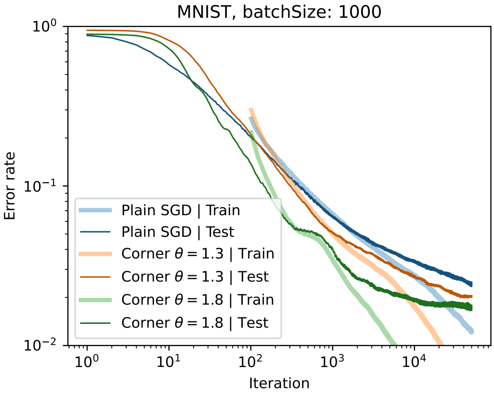
Theory: acceleration $\theta_{\max} \approx 1.35$

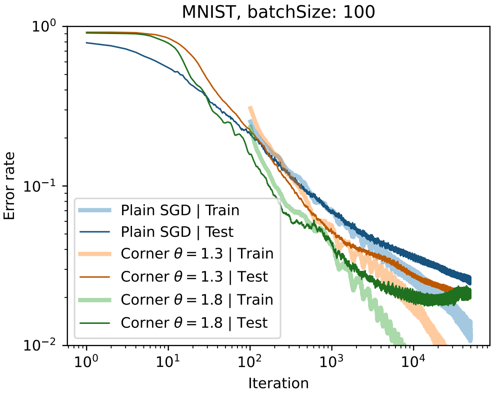



Loss during training, $|B| = 1000$


Loss during training, $|B| = 100$

Train and test accuracy





1. Overview

2. In-depth exposition

The regression problem

Least squares fitting of the **linear** target function $y(\mathbf{x})$ by a linear model:

$$L(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\rho}[(\mathbf{x}^T\mathbf{w} - y(\mathbf{x}))^2],$$

where

- $\mathbf{x} \in \mathcal{H}$ are data points in a Hilbert space $\mathcal{H}$, are described by a distribution ρ
- $\mathbf{w} \in \mathcal{H}$ is the vector of parameters

Assume $\mathbf{w}_* = \arg\min_{\mathbf{w}} L(\mathbf{w})$ is a minimizer, and $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_*$. Then

$$L(\mathbf{w}) = \frac{1}{2}\Delta\mathbf{w}^T\mathbf{H}\Delta\mathbf{w}$$

with the Hessian

$$\mathbf{H} = \mathbb{E}_{\mathbf{x}\sim\rho}[\mathbf{x}\mathbf{x}^T]$$

Assume $\dim(\mathcal{H}) = \infty$ and Hessian $\mathbf{H}$ has a discrete spectrum $\lambda_k \searrow 0$.

The setting is applicable to overparameterized NN's close to linearity (e.g., in the Neural Tangent Kernel regime)

The stochastic mini-batch setting

Mini-batch loss:

$$L_B(\mathbf{w}) = \frac{1}{2|B|}\sum_{m=1}^{|B|}(\mathbf{x}_m^T\mathbf{w} - y(\mathbf{x}_m))^2, \quad B = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|B|}\}$$

The mini-batches B_t at different iterations t are random and independent

Deterministic (full-batch) GD = $\lim_{|B|\to\infty}$ SGD

Stationary generalized SGD with memory M

$$\begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_t \\ \mathbf{u}_{t+1} \end{pmatrix} = \begin{pmatrix} -\alpha & \mathbf{b}^T \\ \mathbf{c} & D \end{pmatrix} \begin{pmatrix} \nabla L_{B_t}(\mathbf{w}_t) \\ \mathbf{u}_t \end{pmatrix}, \quad t = 0, 1, 2, \ldots$$

Here

- $\mathbf{w}_t \in \mathcal{H}$: the main current state vector
- $\mathbf{u}_t \in \mathbb{R}^M \otimes \mathcal{H}$: a set of M "generalized momentum vectors"
- $\alpha \in \mathbb{R}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^M, D \in \mathbb{R}^{M \times M}$: parameters of the algorithm

$M = 0$: plain SGD
$M = 1$: includes Heavy Ball

Equivalent sequential form and characteristic polynomial

Memory-M GD admits an equivalent **sequential form**:

$$\mathbf{w}_{t+M+1} = \sum_{m=0}^{M} p_m \mathbf{w}_{t+m} + \sum_{m=0}^{M} q_m \nabla L(\mathbf{w}_{t+m}), \quad t = 0, 1, \ldots,$$

The coefficients p_m, q_m are found from the **characteristic polynomial**

$$\begin{aligned} \chi(\mu, \lambda) &= \det(\mu - S_\lambda) = P(\mu) - \lambda Q(\mu), \\ P(\mu) &= \mu^{M+1} - \sum_{m=0}^{M} p_m \mu^m, \\ Q(\mu) &= \sum_{m=0}^{M} q_m \mu^m, \end{aligned}$$

where S_λ is the noiseless one-step transition matrix in the λ-eigenspace of $\mathbf{H}$:

$$\begin{pmatrix} \Delta \mathbf{w}_{t+1} \\ \mathbf{u}_{t+1} \end{pmatrix} = S_\lambda \begin{pmatrix} \Delta \mathbf{w}_t \\ \mathbf{u}_t \end{pmatrix}, \quad S_\lambda = \begin{pmatrix} 1 & \mathbf{b}^T \\ 0 & D \end{pmatrix} + \lambda \begin{pmatrix} -\alpha \\ \mathbf{c} \end{pmatrix} (1, \mathbf{0}^T)$$

Spectrally-Expressible approximation

We want to study $L_t = \mathbb{E}_{B_1,\ldots,B_t} L(\mathbf{w}_t)$ – averaged, deterministic loss at step t

But computation of L_t involves 4'th moments of data distribution ρ and so generally requires more information than the spectral properties of $\mathbf{H}$ and of solution $\mathbf{w}_*$

Spectrally-Expressible (SE) approximation:

$$\mathbb{E}_{\mathbf{x}\sim\rho}[\mathbf{x}\mathbf{x}^T\mathbf{C}\mathbf{x}\mathbf{x}^T] \approx \tau_1 \operatorname{Tr}[\mathbf{HC}]\mathbf{H} - (\tau_2 - 1)\mathbf{HCH}$$

- Holds exactly with some τ_1, τ_2 for some natural classes of data (translation-invariant, Gaussian)
- Holds approximately in MNIST and some other real world data
- If holds only as inequality, can still be used to upper/lower bound true L_t

We will use the SE approximation with $\tau_2 = 0$ (simplifies computations)

Propagator expansion of the loss

$$L_t \equiv \mathbb{E}L(\mathbf{w}_t) = \frac{1}{2}\Big(V_{t+1} + \sum_{m=1}^{t} \sum_{0<t_1<\ldots<t_m<t+1} U_{t+1-t_m} U_{t_m-t_{m-1}} U_{t_{m-1}-t_{m-2}} \cdots U_{t_2-t_1} V_{t_1}\Big)$$

with

- **Signal propagators**

$$V_t = \sum_{\lambda_k \in \mathrm{spec}(H)} \lambda_k (\mathbf{w}_*^T \mathbf{e}_k)^2 \Big|\begin{pmatrix} 1 & 0 \end{pmatrix} S_{\lambda_k}^{t-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}\Big|^2$$

- **Noise propagators**

$$U_t = \frac{\tau_1}{|B|} \sum_{\lambda_k \in \mathrm{spec}(H)} \lambda_k^2 \Big|\begin{pmatrix} 1 & 0 \end{pmatrix} S_{\lambda_k}^{t-1} \begin{pmatrix} -\alpha \\ \mathbf{c} \end{pmatrix}\Big|^2$$

Deterministic GD: $U_t \equiv 0$ and $L_t = \frac{1}{2} V_{t+1}$

Convergence of SGD = Convergence of GD & $U_\Sigma < 1$

Consider **total noise coefficient** $U_\Sigma = \sum_{t\geq 1} U_t$

Theorem.

1. **[Convergence]** Suppose that $U_\Sigma < 1$. If V_t is bounded (resp., $V_t \to 0$), then also L_t is bounded (resp., $L_t \to 0$).
2. **[Divergence]** If $U_\Sigma > 1$ and $V_t > 0$ for at least one t, then $\sup_{t=1,2,\dots} L_t = \infty$.

The convergent phases: signal- and noise-dominated

Theorem. Assume $V_t = C_V t^{-\xi_V}(1 + o(1))$, $U_t = C_U t^{-\xi_U}(1 + o(1))$, and $U_\Sigma < 1$.

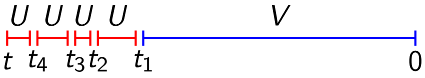
1. **[Signal-dominated regime]** If $\xi_U > \xi_V$, then

$$L_t = \frac{C_V}{2(1 - U_\Sigma)} t^{-\xi_V}(1 + o(1)).$$

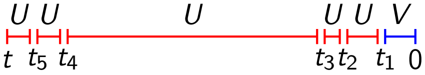
2. **[Noise-dominated regime]** If $\xi_V > \xi_U$, then

$$L_t = \frac{V_\Sigma C_U}{2(1 - U_\Sigma)^2} t^{-\xi_U}(1 + o(1)).$$

Signal-dominated



Noise-dominated



Stability and effective learning rate

$S_{\lambda=0} = \left(\begin{smallmatrix} 1 & \mathbf{b}^T \\ 0 & D \end{smallmatrix}\right)$ has eigenvalue 1. Assume it's the largest eigenvalue.

For stability, we need the respective eigenvalue μ_λ of S_λ to decrease as λ increases from 0

Theorem.

$$\mu_\lambda = 1 - \alpha_{\text{eff}}\lambda + O(\lambda^2), \quad \lambda \searrow 0$$

with the **effective learning rate**

$$\alpha_{\text{eff}} = -\frac{Q(1)}{\frac{d}{d\mu}P(1)}$$

Power-law phase diagram for SGD with **finite** memory

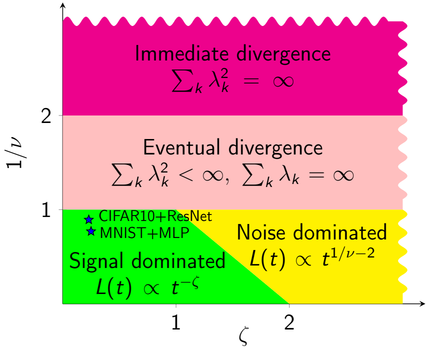
Power-law spectral assumptions:

$$\lambda_k = \Lambda k^{-\nu}(1+o(1)), \quad k \to \infty \qquad \text{(eigenvalue decay)}$$

$$\sum_{k:\lambda_k<\lambda} \lambda_k(\mathbf{w}_*^T \mathbf{e}_k)^2 = Q\lambda^{\zeta}(1+o(1)), \quad \lambda \to 0 \qquad \text{(source condition)}$$

Theorem.

1. **[Divergent phase]** If $\nu < 1$, then $\sup_t L_t = \infty$
2. **[Signal-dominated phase]** If $\zeta < 2 - 1/\nu$, then
$$L_t = \frac{\alpha_{\text{eff}}^{-\zeta}}{1-U_\Sigma} Q\Gamma(\zeta+1)2^{-\zeta-1}(1+o(1))t^{-\zeta}$$
3. **[Noise-dominated phase]** If $2 - 1/\nu < \zeta$, then
$$L_t = \frac{\alpha_{\text{eff}}^{1/\nu} V_\Sigma}{|B|(1-U_\Sigma)^2} \frac{\Lambda^{1/\nu}\Gamma(2-1/\nu)}{\nu 2^{3-1/\nu}}(1+o(1))t^{\frac{1}{\nu}-2}.$$



In signal-dominated regime: accelerate = increase α_{eff} while keeping $U_\Sigma < 1$

Contour representations of propagators

Recall the characteristic polynomials $\chi(\mu, \lambda) = \det(\mu - S_\lambda) = P(\mu) - \lambda Q(\mu)$. Define

$$\boxed{\Psi(\mu) = \frac{P(\mu)}{Q(\mu)}}$$

Sequential form of GD in the frequency domain obeys $\widehat{\nabla L}(\mu) = \Psi(\mu)\widehat{\mathbf{w}}(\mu)$ with $|\mu| = 1$
$\implies \Psi$ can be interpreted as **frequency response function**

The signal and noise propagators, and hence the loss L_t, are **completely determined by the function Ψ**:

$$U_t = \frac{\tau_1}{|B|} \sum_{k=1}^{\infty} \lambda_k^2 \Big| \frac{1}{2\pi i} \oint_{|\mu|=1} \frac{\mu^{t-1} d\mu}{\Psi(\mu) - \lambda_k} \Big|^2,$$

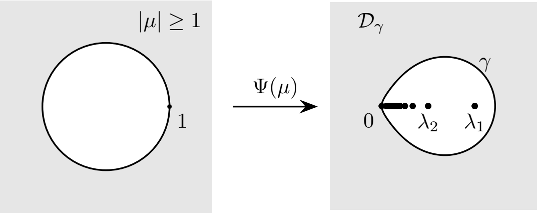
$$V_t = \sum_{k=1}^{\infty} \lambda_k (\mathbf{e}_k^T \mathbf{w}_*)^2 \Big| \frac{1}{2\pi i} \oint_{|\mu|=1} \frac{\mu^{t-1} \Psi(\mu) d\mu}{(\Psi(\mu) - \lambda_k)(\mu - 1)} \Big|^2$$

The function Ψ: properties and identification with contours

In memory-M SGD $\deg(P) = M + 1$ and $\deg(Q) \leq M$, so Ψ is rational and $\lim_{\mu\to\infty} \Psi(\mu) = \infty$

Stability: S_λ does not have eigenvalues μ with $|\mu| > 1 \iff \mathrm{spec}(\mathbf{H}) \subset \mathbb{C} \setminus \Psi(\{|\mu| \geq 1\})$

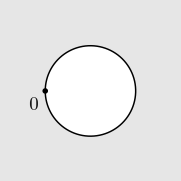
Assuming injectivity of Ψ on $|\mu| \geq 1$, the map Ψ can be recovered from the contour $\gamma = \Psi(\{|\mu| = 1\})$ by the Riemann mapping theorem



Examples of contours $\gamma = \Psi(\{|\mu| = 1\})$

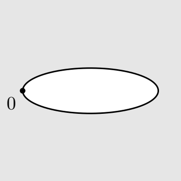
- **Plain GD**: $\Psi(\mu) = -\frac{\mu-1}{\alpha}$

  A circle



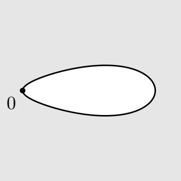
- **Heavy Ball**: $\Psi(\mu) = -\frac{(\mu-1)(\mu-\beta)}{\alpha\mu}$

  An ellipse with eccentricity $\frac{2\sqrt{\beta}}{1+\beta}$

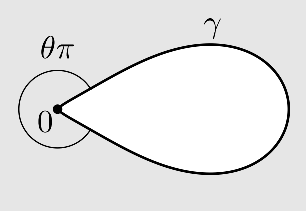


- **General memory-1**: $\Psi(\mu) = \frac{(\mu-1)(\mu-\beta)}{q_0+q_1\mu}$

  A Zhukovsky airfoil (a degree-4 algebraic set)



Corner algorithms



$\Psi(\mu) = -c_\Psi(\mu{-}1)^\theta(1{+}o(1))$ as $\mu \to 1, \quad c_\Psi > 0, \quad 1 < \theta < 2$

Ψ is **irrational**, requires memory $M = \infty$

Intuition: $\alpha_{\mathrm{eff}} = -(\frac{d}{d\mu}\Psi(1+))^{-1} = +\infty$

Theorem.

1. **(Noise propagators)** $\boxed{U_t = C_U t^{\theta/\nu-2}(1+o(1))}$, with the coefficient

$$C_U = \frac{\tau_1}{|B|}\Lambda^{1/\nu}\int_\infty^0 r^2 F_U^2(r) dr^{-\theta/\nu} < \infty, \quad F_U(r) = \frac{1}{2\pi i}\int_{i\mathbb{R}} \frac{e^{rz}dz}{c_\Psi z^\theta + 1}.$$

2. **(Signal propagators)** $\boxed{V_t = C_V t^{-\theta\zeta}(1+o(1))}$, with the coefficient

$$C_V = Q\int_0^\infty F_V^2(r) dr^{\theta\zeta} < \infty, \quad F_V(r) = \frac{1}{2\pi i}\int_{i\mathbb{R}} \frac{c_\Psi z^{\theta-1}e^{rz}dz}{c_\Psi z^\theta + 1}.$$

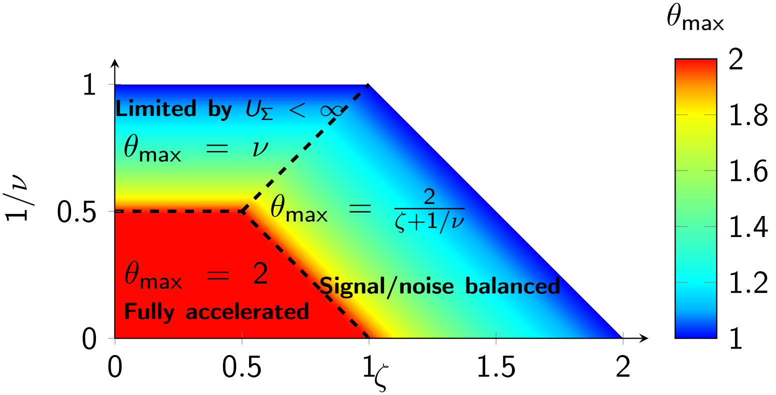
Increasing θ **improves signal propagators, but degrades noise propagators**

Acceleration phase diagram in the signal regime

Theorem. Let $\theta_{\max}$ denote the supremum of those θ for which there exists a corner algorithm and batch size B such that $L_t = O(t^{-\theta\zeta})$. Then

$$\theta_{\max} = \min\left(2, \nu, \frac{2}{\zeta + 1/\nu}\right)$$

Three sub-phases:


$\theta_{\max}$
Limited by $U_\Sigma < \infty$
$\theta_{\max} = \nu$
$\theta_{\max} = \frac{2}{\zeta+1/\nu}$
Signal/noise balanced
$\theta_{\max} = 2$
Fully accelerated
$1/\nu$
ζ