Large Spikes in SGD: A Large-Deviations View of Catapults

Benjamin Gess
TU Berlin & MPI MiS Leipzig

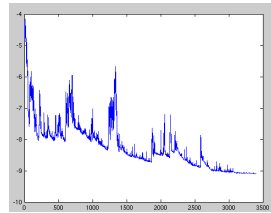
SIAM Conference on Optimization (OP26), Edinburgh, June 2026

Based on joint work with D. Heydecker

Goals: (1) Explain how the deterministic *catapult phase* refines, through minibatching noise, into *guaranteed vs polynomially-likely spikes*

(2) Rigorous justification

(3) Occurence of multiple spikes



Roadmap

1. Recap: Full-batch GD - deterministic catapults at large learning rates, NTK.
   (Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., & Gur-Ari, G. (2020). The large learning rate phase of deep learning: the catapult mechanism. arXiv:2003.02218)
2. Recap: Empirics for catapults in SGD
   (Zhu, L., Liu, C., Radhakrishnan, A., & Belkin, M. (ICML 2024). Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning)
3. Main findings:
   1. internal structure of SGD catapult region.
   2. Catapults and curvature decrease
4. Intuition: multiplicative random walk + Cramér exponent.

Recap: NTK

Given network function $F : \mathbb{R}^{2n} \times \mathbb{R}^d \to \mathbb{R}$, mean square loss

$$L(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(F(\Theta; s_i) - y_i\right)^2 .$$

The NTK $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined by

$$K(s, s') := \frac{1}{m} \sum_{\mu=1}^{2n} \frac{\partial F(\Theta; s)}{\partial \Theta_\mu} \frac{\partial F(\Theta; s')}{\partial \Theta_\mu}.$$

with maximum eigenvalue λ.

[Jacot, Gabriel, Hongler; 2018]: For fixed time-span, in infinite width, zero learning rate limit, limit dynamics are well approximated by quadratic minimization with Hessian K.

Recap: NTK

Model case, consider one-hidden-layer linear network

$$F(\Theta; s) = \frac{1}{\sqrt{n}} \sum_{r=1}^{n} a_r \, \phi(w_r s), \qquad \Theta = ((w_r, a_r))_{r=1}^{n}.$$

Here n is the width (number of neurons in the hidden layer), $a \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$ are the model parameters (collectively denoted Θ), and $s \in \mathbb{R}^d$ is the data. At initialization, the weights are drawn from $\mathcal{N}(0, 1)$.

One-sample, one-hidden-layer linear network ($s = 1, y = 0$) with linear activation $\phi(w) = w$: Get

$$\mu = \frac{1}{\sqrt{n}} a^\top w, \quad L = \tfrac{1}{2}\mu^2, \quad \lambda(t) = \frac{1}{n}(\|a(t)\|_2^2 + \|w(t)\|_2^2).$$

Recap: NTK

Full-batch GD yields:

$$\mu(t+1) = \left(1 - \eta\lambda(t) + \eta\frac{\mu^2(t)}{n}\right)\mu(t),$$
$$\lambda(t+1) = \lambda(t) + \eta\frac{\mu^2(t)}{n}(\eta\lambda(t) - 4).$$

Observe: NKT limit

$$\mu(t+1) = (1 - \eta\lambda(t))\,\mu(t),$$
$$\lambda(t+1) = \lambda(t).$$

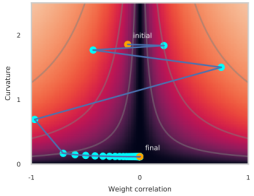
Threshold: $\eta_{\text{crit}} = 2/\lambda_0$.
But also note threshold: $\eta_{\text{max}} = 4/\lambda_0$.

Recap: "catapult mechanism"

Recall

$$\mu(t+1) = \left(1 - \eta\lambda(t) + \eta\frac{\mu^2(t)}{n}\right)\mu(t),$$

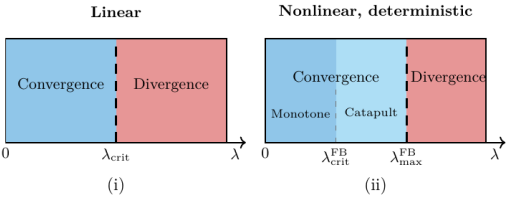
$$\lambda(t+1) = \lambda(t) + \eta\frac{\mu^2(t)}{n}(\eta\lambda(t) - 4).$$



(Source: Lewkowycz et. al. 2020)

Training wide networks with MSE exhibits learning-rate phases determined by the initial curvature scale (top NTK eigenvalue) λ_0:

- **Lazy / NTK phase:** $\eta < 2/\lambda_0$ (linearized dynamics stable).
- **Catapult phase:** $2/\lambda_0 < \eta < \eta_{\max}$ (loss can rise then converge; curvature collapses).
- **Divergent phase:** $\eta > \eta_{\max}$.

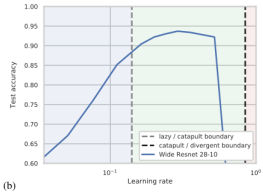


Recap: Catapult mechanism and sharpness

Recall

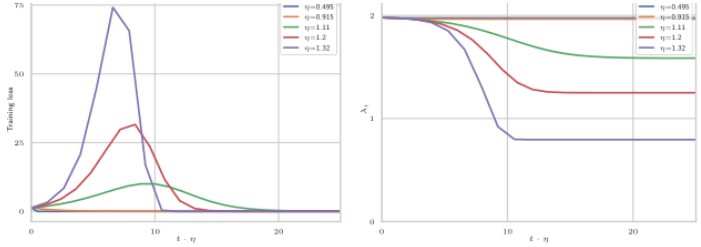
$$\mu(t+1) = \left(1 - \eta\lambda(t) + \eta\frac{\mu^2(t)}{n}\right)\mu(t),$$

$$\lambda(t+1) = \lambda(t) + \eta\frac{\mu^2(t)}{n}(\eta\lambda(t) - 4).$$

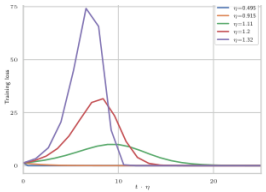


(Source: Lewkowycz et. al. 2020)

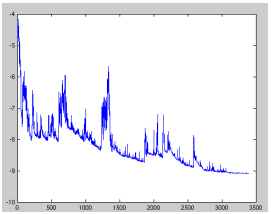
Observe: During spikes sharpness decreases. Large LR can yield a transient loss increase and then better solutions inside a narrow stable window.



Empirical Observation: Spikes in SGD



(Source: Lewkowycz et. al. 2020)



(Source:wikipedia)

Question: Why does SGD exhibit repeated sharp spikes in training loss, while GD shows only a single catapult?Smaller batch $\rightarrow$ higher variance of NTK eigenvalues $\rightarrow$ more frequent violations of

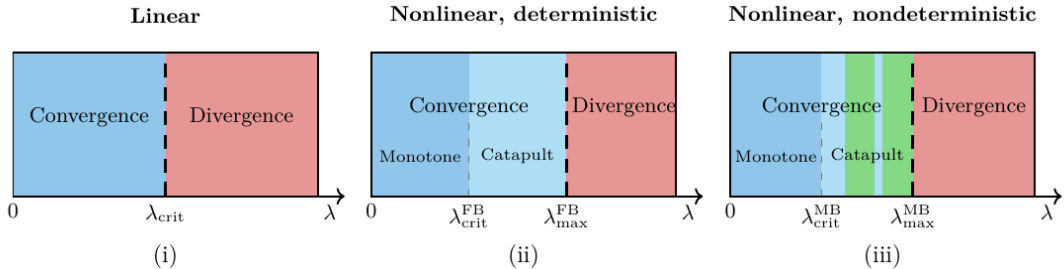
$$\eta < \eta_{\mathrm{crit}}(s_{\mathsf{batch}})$$

Hence:

$$\text{Small batch} \Rightarrow \text{More catapults}$$

Empirically confirmed across FCNs, CNNs, WideResNets, ViTs. (Zhu, Liu, Radhakrishnan, Belkin, 2023)

Internal structure of SGD catapult region.



Green ('inflationary') and pale blue ('deflationary') stripes in (iii).
In general, the critical and maximal curvatures $\lambda^{\text{MB}}_{\text{crit,max}}$ for minibatching are strictly smaller than their fullbatch counterparts $\lambda^{\text{FB}}_{\text{crit,max}}$.

Refinement: in the nonlinear, *nondeterministic* (SGD) case, the catapult region has *internal structure* (inflationary vs deflationary sub-regimes).

Model and two scalar state variables (linear activation case)

Shallow NTK-scaled model:

$$F(\Theta; s) = \frac{1}{\sqrt{n}} \sum_{r=1}^{n} a_r \, \phi(w_r s), \qquad \Theta = ((w_r, a_r))_{r=1}^{n}.$$

Dataset $\{(s_i, 0)\}_{i=1}^{m}$ sampled i.i.d. with probabilities $\{p_i\}$ (minibatch size $b = 1$).

For **linear activation** $\phi(w) = w$, define

$$\mu(t) = \frac{1}{\sqrt{n}} \sum_{r=1}^{n} a_r(t) w_r(t), \qquad \lambda(t) = \frac{1}{n} \sum_{r=1}^{n} (a_r(t)^2 + w_r(t)^2).$$

Think: loss $\ell(t) \propto \mu(t)^2$, curvature/NTK scale $\sim \lambda(t)$. Get

$$\mu(t+1) = \left(1 - \eta \lambda(t) s_{i(t+1)}^2 + \frac{\eta^2 s_{i(t+1)}^4 \mu(t)^2}{n}\right) \mu(t);$$

$$\lambda(t+1) = \lambda(t) + \frac{\mu(t)^2 \eta}{n} (\eta \lambda(t) s_{i(t+1)}^4 - 4 s_{i(t+1)}^2).$$

If $\frac{\eta^2 s_{i(t+1)}^4 \mu(t)^2}{n}$, $\frac{\mu(t)^2 \eta}{n} << 1$ then enter linear regime.

Stochastic GD: If

$$\frac{\eta^2 s_{i(t+1)}^4 \mu(t)^2}{n}, \frac{\mu(t)^2 \eta}{n} << 1$$

then

$$\mu(t+1) = \left(1 - \eta\lambda(t) s_{i(t+1)}^2\right) \mu(t);$$
$$\lambda(t+1) = \lambda(t).$$

and $\mu(t)$ approximately evolves by a **multiplicative random walk**:

$$\begin{aligned}
\mu(t+1) &\approx \left(1 - \eta\lambda(0) s_{i(t+1)}^2\right) \mu(t) \\
&= \mu_0 \prod_{u=1}^{t} \left(1 - \eta\lambda_0 s_{i(u)}^2\right) \\
\log|\mu(t)| &\approx \log|\mu_0| + \sum_{u=1}^{t} \log|1 - \eta\lambda_0 s_{i(u)}^2| \\
&\approx_{LLN} \log|\mu_0| + t \underbrace{\mathbb{E}_s \log|1 - \eta\lambda_0 s^2|}_{=:G(\lambda_0)}
\end{aligned}$$

Full batch GD: If

$$\frac{\eta^2 (\mathbb{E}_s s^2)^2 \mu(t)^2}{n}, \frac{\mu(t)^2 \eta}{n} << 1$$

then

$$\mu(t+1) = \left(1 - \eta\lambda(t) \mathbb{E}_s s^2\right) \mu(t);$$
$$\lambda(t+1) = \lambda(t).$$

and

$$\begin{aligned}
\mu(t) &\approx \mu_0 \prod_{u=1}^{t} \left(1 - \eta\lambda_0 \mathbb{E}_s s^2\right) \\
&= \mu_0 \left(1 - \eta\lambda_0 \mathbb{E}_s s^2\right)^t \\
\log|\mu(t)| &\approx \log|\mu_0| + t \underbrace{\log|1 - \eta\lambda_0 \mathbb{E}_s s^2|}_{=:G_{FB}(\lambda_0)}
\end{aligned}$$

Related dynamical system perspective: [Chemnitz, Engel, JMLR; 2025]

Stochastic GD:

$$\log|\mu(t)| \approx \log|\mu_0| + t\underbrace{\mathbb{E}_s \log|1-\eta\lambda_0 s^2|}_{=:G(\lambda_0)}$$

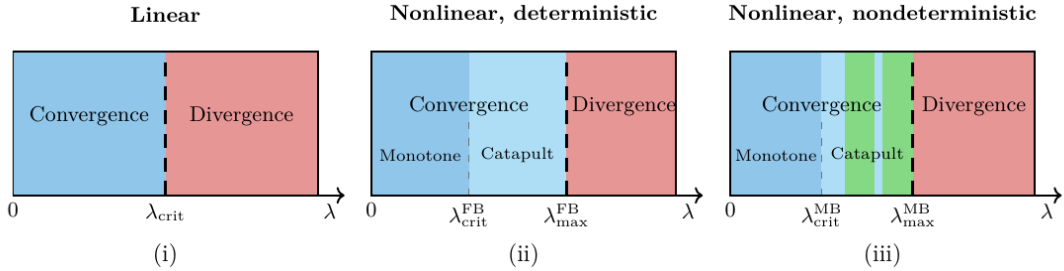
Full batch GD:

$$\log|\mu(t)| \approx \log|\mu_0| + t\underbrace{\log|1-\eta\lambda_0 \mathbb{E}_s s^2|}_{=:G_{FB}(\lambda_0)}$$

Spikes correspond to $|\mu(t)|$ getting large.

- **Inflationary:** $G(\lambda_0) > 0 \Rightarrow$ spikes are typical/guaranteed.
- **Deflationary:** $G(\lambda_0) < 0 \Rightarrow$ spikes are large-deviation events.

Since $\log|\cdot|$ is neither convex nor concave, G and G_{FB} are not comparable.



Example: $G(\lambda)$ can be non-monotone (sign changes)

Consider datapoints

$$\text{i)} \quad \{(s_i, p_i)\} = \{(1, 0.5), (1.3, 0.5)\}$$
$$\text{ii)} \quad \{(s_i, p_i)\} = \{(1, 0.83), (\sqrt{2}, 0.17)\}$$

and recall

$$G(\lambda) := \sum_{i=1}^{m} p_i \log \left|1 - \eta\lambda s_i^2\right|.$$

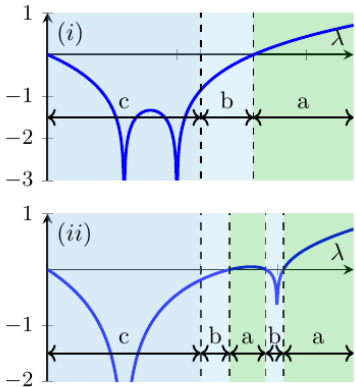


FIGURE 2. Plots of $G(\lambda)$ for the examples (1.17 - 1.18)

Even for two datapoints, $G(\lambda)$ can change sign multiple times: increasing curvature can make spikes *more* or *less* likely (non-monotonicity).

Deflationary regime $G(\lambda_0) < 0$: Large deviations and polynomial spike probability.

Define the **Cramér/LDP exponent**

$$\vartheta(\lambda) := \text{the unique positive root of } \sum_i p_i |1 - \eta\lambda s_i^2|^\theta = 1.$$

Recall: **Deflationary:** $G(\lambda_0) < 0 \Rightarrow$ spikes are rare events, but only *polynomially* rare:

Theorem

In particular, at the medium **spike scale** $M \sim \sqrt{n/\eta}$*: (up to polylogs)*

$$\mathbb{P}(\textit{medium large spike}) \approx (n/\eta)^{-\vartheta(\lambda_0)/2}.$$

Linear regime: $\frac{\eta^2 s_{i(t+1)}^4 \mu(t)^2}{n}, \frac{\mu(t)^2 \eta}{n} << 1$

Plot: $\vartheta(\lambda)$ and the resulting probability scale

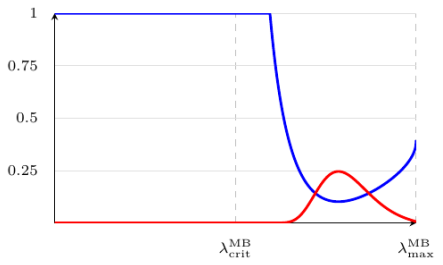


FIGURE 3. $\vartheta(\lambda)$ (blue) and $n^{-\vartheta(\lambda)/2}$ with $n = 10^{12}$ (red) for the dataset (1.26).

Large Spikes and curvature decrease

Assume:

- $\lambda_{MB}^{\text{crit}} < \lambda_0 < \lambda_{MB}^{\max}$
- Moderate spike has reached scale: $M \lesssim \frac{1}{\log^{1/\beta}(n/\eta)} \sqrt{\frac{n}{\eta}}$
- Let $\vartheta(\lambda_0) = \sup\left\{\theta \geq 0 : \sum_{i=1}^{m} p_i \left|1 - \lambda_0 s_i^2\right|^\theta \leq 1\right\} > 0.$

Theorem 5 (Large Spikes)

Conditional on reaching a moderate spike, the probability of a *large spike* that reduces curvature from λ_0 to λ decays at most polynomially:

$$\mathbb{P}(\text{large spike reducing } \lambda_0 \to \lambda) \gtrsim \left(\frac{\sqrt{n/\eta}}{|\mu_0|}\right)^{-\vartheta(\lambda_0)} (\lambda_0 - \lambda)^\alpha$$

Why spikes trigger curvature reduction (the catapult mechanism, stochastic edition)

At spike heights $|\mu(t)| \sim \sqrt{n/\eta}$, curvature updates become $O(1)$:

$$\lambda(t+1) - \lambda(t) = \frac{\eta}{n}\mu(t)^2\big(\eta\lambda(t)s_{i(t+1)}^4 - 4s_{i(t+1)}^2\big).$$

In the minibatch catapult window $\lambda(t) < 4/(\eta s_\star^2)$, for all i,

$$\eta\lambda(t)s_i^4 - 4s_i^2 \le s_i^2(\eta\lambda(t)s_\star^2 - 4) < 0,$$

so a large spike forces $\lambda(t)$ down.

Message: the same geometric story (instability $\Rightarrow$ spike $\Rightarrow$ curvature collapse) persists, but the spike itself may be typical or LDP-rare.

However: In principle, sharpness could also gradually decrease

Main Technical Challenges & Their Resolution

Challenge 1: Nonlinear Coupling - The updates for $(\mu(t), \lambda(t))$ are coupled:

$$\mu(t+1) = \Big(1 - \eta\lambda(t)s_i^2 + O(\mu(t)^2/n)\Big)\mu(t),$$

$$\lambda(t+1) = \lambda(t) - \frac{\eta}{n}|\mu(t)|^2 + \dots$$

Resolution**:** Work in a stopping region $|\mu| \ll \sqrt{n/\eta}$, $\lambda \approx \lambda_0$, where the system is a perturbation of a multiplicative random walk. Carefully control variations in I.

Challenge 2: Large deviations estimates - The drift depends on $\lambda(t)$, which evolves.

Resolution**:** Construct sub- and supermartingales:

$$|\mu(t)|^{\vartheta}$$

and use optional stopping + change of measure (Cramer tilting) to obtain polynomial decay.

Challenge 3: Excluding Slow Kernel Drift Could curvature decrease via many small fluctuations?

Resolution**:** Use Khasminskii-type exponential moment bounds to show slow escape has probability

$$\leq \exp\Big(-c(n/\eta M^2)^{\beta/2}\Big).$$

Summary: catapults in SGD

1. Inside the catapult-capable region, the drift

$$G(\lambda_0) = \sum_i p_i \log|1 - \eta\lambda_0 s_i^2|$$

   splits behavior into inflationary (spikes typical) vs deflationary (spikes LDP-rare).
2. Deflationary spikes remain *polynomially likely*: (large spike) $\approx (n/\eta)^{-\vartheta(\lambda_0)/2}$.
3. Significant curvature reduction without spikes is exponentially unlikely: spikes are the dominant escape mechanism.

Extension: ReLu activation.

References

S. Gess, M. Heydecker, *Large spikes in SGD: a large-deviations view of catapults*, arXiv, 2026.