Generalization of Stochastic Gradient Methods: From Heavy Tails to Entropy Flows

Benjamin Dupuis

Liverpool Discrete Mathematics Colloquium - 2026

About me

- Postdoc in the department of statistics at the University of Oxford
- Previously a PhD student in the SIERRA team at INRIA and ENS

Research interests

- Generalization bounds
- Stochastic optimization
- Diffusion models
- Differential privacy
- ...

About me

- Postdoc in the department of statistics at the University of Oxford
- Previously a PhD student in the SIERRA team at INRIA and ENS

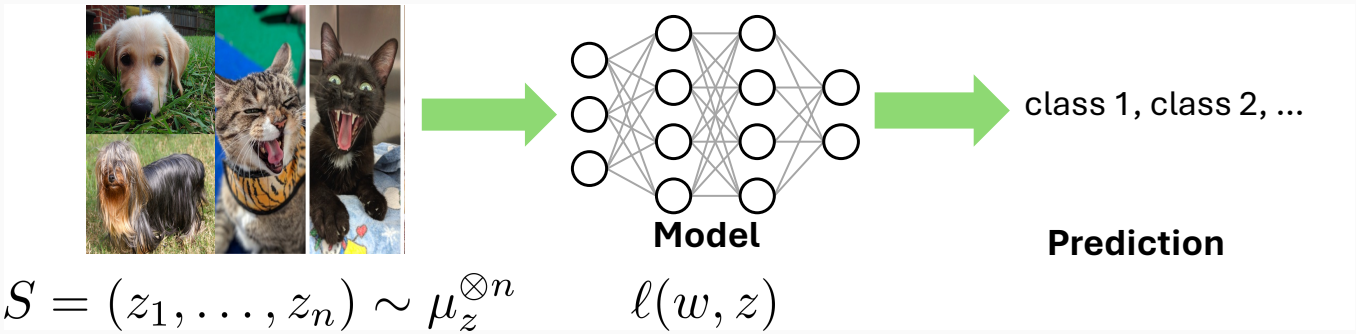
Research interests

- Generalization bounds
- Stochastic optimization
- Diffusion models
- Differential privacy
- ...

Today's talk

- Quick overview of some generalization bounds for SGD
- Our work: Heavy tails and extension of the entropy flow method
- Joint work with Umut Şimşekli, Maxime Haddouche and George Deligiannidis

Supervised Learning Setup



- Data distribution μ_z over a data space $\mathcal{Z}$
- A model (or loss) $\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$
- Goal: minimize the population risk

$$\min_{w \in \mathbb{R}^d} \left\{ \mathcal{R}(w) := \int \ell(w, z) \mathrm{d}\mu_z(z) \right\}$$

Supervised learning setup

- In practice, we use a *dataset* $S = (z_1, \ldots, z_n) \sim \mu_z^{\otimes n}$.
- The minimized quantity is the empirical risk:

$$\mathcal{R}(w) \Longrightarrow \widehat{\mathcal{R}}_S(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$$

- highly non-convex problem
- Generalization error:

$$\boxed{G_S(W_S) := \mathcal{R}(W_S) - \widehat{\mathcal{R}}_S(W_S)}$$

- Learning algorithm: $\mathscr{A} : S \mapsto W_S \in \mathbb{R}^d$.

Stochastic Gradient Descent (SGD)

- Gradient descent: $W_{k+1}^S = W_k^S - \eta \nabla \widehat{\mathcal{R}}_S(W_k^S)$.
- Not computationally scalable for modern models and datasets

Stochastic Gradient Descent (SGD)

- Gradient descent: $W_{k+1}^S = W_k^S - \eta \nabla \widehat{\mathcal{R}}_S(W_k^S)$.
- Not computationally scalable for modern models and datasets
- Stochastic Gradient Descent (SGD):

$$\boxed{W_{k+1}^S = W_k^S - \frac{\eta}{b} \sum_{i \in \Omega_k} \nabla \ell(W_k^S, z_i),}$$

  where η is the learning rate and $b := |\Omega_k|$ is the batch size.
- More generally:

$$W_{k+1}^S = W_k^S - \eta \nabla \widehat{\mathcal{R}}_S(W_k^S) + \zeta_k,$$

  with i.i.d. zero-mean noise ζ_k.
- Many variants have been proposed

Generalization Error & SGD

- Modern models can easily overfit the training dataset [1]
- Classical uniform bounds [1] cannot explain the empirical success
- SGD has an implicit bias towards well-generalizing solutions

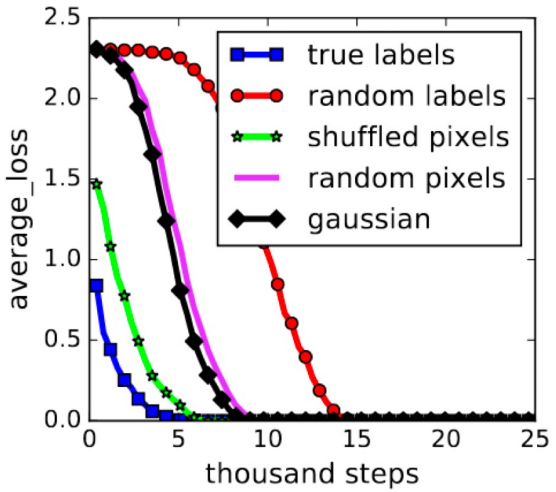


Fig. 1. Overparameterized models can feat random labels

Figure: *MLP on CIFAR10 with random labels [22]*

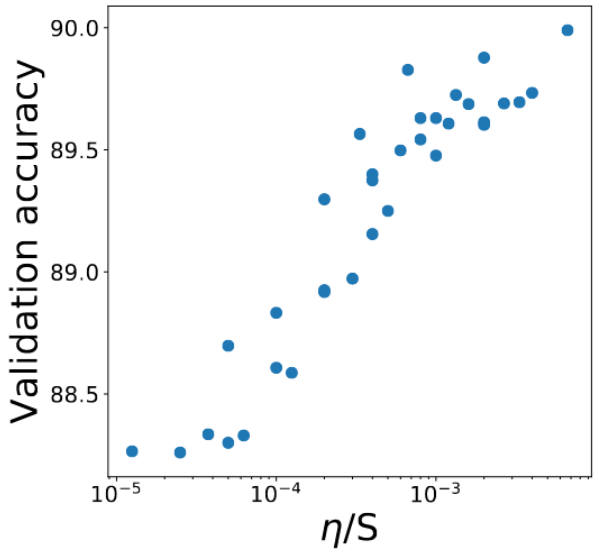


Figure: *Generalization vs (learning rate) / (batch size). MLP on FashionMNIST [12]*

- Importance of the noise structure and hyperparameters

[1] Zhang et al, Understanding Deep Learning Requires Rethinking Generalization, 2017.

Algorithm-dependent generalization bounds

There is only a few types of methods:

- **Information-theoretic bounds:** Mutual information[2], PAC-Bayes
  - Statistical dependence between W_S and S

$$\mathbb{E}\left[G_S(W_S)\right] \lesssim \sqrt{\frac{\mathrm{I}_1(W_S, S)}{n}}$$

[2] Information-Theoretic Analysis of Generalization Capability of Learning Algorithms, Aolin Xu and Maxim Raginsky, NeurIPS 2017.

[3] Stability and generalization, Olivier Bousquet and André Elissef, JMLR 2002.

Algorithm-dependent generalization bounds

There is only a few types of methods:

- **Information-theoretic bounds:** Mutual information[2], PAC-Bayes
  - Statistical dependence between W_S and S

$$\mathbb{E}\left[G_S(W_S)\right] \lesssim \sqrt{\frac{\mathrm{I}_1(W_S, S)}{n}}$$

- **Algorithmic stability**[3]
  - Sensitivity to a perturbation of the dataset

$$\mathbb{E}\left[G_S(W_S)\right] \leq \sup_{S \simeq S'} \sup_{z \in \mathcal{Z}} \mathbb{E}_U\left[\ell(W_S, z) - \ell(W_{S'}, z)\right]$$

---

[2] Information-Theoretic Analysis of Generalization Capability of Learning Algorithms, Aolin Xu and Maxim Raginsky, NeurIPS 2017.

[3] Stability and generalization, Olivier Bousquet and André Elissef, JMLR 2002.

Algorithm-dependent generalization bounds

There is only a few types of methods:

- **Information-theoretic bounds:** Mutual information[2], PAC-Bayes
  - Statistical dependence between W_S and S

$$\mathbb{E}\left[G_S(W_S)\right] \lesssim \sqrt{\frac{\mathrm{I}_1(W_S, S)}{n}}$$

- **Algorithmic stability**[3]
  - Sensitivity to a perturbation of the dataset

$$\mathbb{E}\left[G_S(W_S)\right] \leq \sup_{S \simeq S'} \sup_{z \in \mathcal{Z}} \mathbb{E}_U\left[\ell(W_S, z) - \ell(W_{S'}, z)\right]$$

- **Uniform bounds**: not algorithm-dependent in general. If $W_S \in \mathcal{W}$

$$G_S(W_S) \leq \sup_{w \in \mathcal{W}} G_S(w)$$

---

[2] Information-Theoretic Analysis of Generalization Capability of Learning Algorithms, Aolin Xu and Maxim Raginsky, NeurIPS 2017.

[3] Stability and generalization, Olivier Bousquet and André Elissef, JMLR 2002.

Overview of existing bounds for mini-batch SGD

Paper	Method	$\ell(w,z)$	Asmp. on η_k	Bound ($\mathcal{O}$)
[16]	stability	L., ∇-M., C.	$\eta_k \leq \frac{2}{M}$	$\frac{L^2}{n}\sum_k \eta_k$
[16]	stability	∇-M., μ-SC.	$\eta_k = \eta \leq \frac{1}{M}$	$\frac{L^2}{\mu n}$
[16]	stability	B., L., ∇-M.	$\eta_k \leq \frac{c}{k}$	$\frac{L^{\frac{2}{\beta c+1}}}{n} T^{\frac{\beta c}{\beta c+1}}$
[18]	stability	C.	$\eta_k = \frac{c}{\sqrt{k}}$	$\frac{T}{\sqrt{n}}\mathrm{Var}(\nabla\ell)^{1/2}$
[18]	stability	L. $\nabla^2\ell$ ρ-L.	$\eta_k = \frac{c}{k}$	see [18, Theorem 4]
[12]	stability	C., L., ∇-L.	$\eta_k = \mathcal{O}(1)$	see [12, Theorem 4.5]
[2]	stability	L.	any	$L^2\left(\sqrt{\sum_{k=1}^T \eta_k^2} + \frac{1}{n}\sum_{k=1}^T \eta_k\right)$
[22]	IT	∇-M., SG.	$\eta_k = \eta$	See [22, Corollary 2]
[29]	stability	p.L., SC.	$\eta_k = \eta$	See [29, Theorem 8]
[29]	stability	p.L., D.	$\eta_k = \eta$	See [29, Theorem 12]
[4]	IT	SG., ∇-M.	$\eta_k \leq \frac{2}{M}$	$\sqrt{\frac{1}{n}\sum_{k=1}^T \eta_k \widehat{\|\nabla^2\ell\|}}$

Table: L. Lipschitz, ∇-M. M-smooth, C. convex, B. bounded, SG. subgaussian, D. dissipatif, p.L. pseudo-Lipschitz, SC. strongly convex

What method for (noisy) SGD?

These methods have advantages and drawbacks:

- **Information-theoretic bounds:** Mutual information, PAC-Bayes
  - Mostly adapted to noisy variants, *e.g.*, SGLD (stochastic gradient Langevin dynamics)

$$W_{k+1}^S = W_k^S - \frac{\eta}{b} \sum_{i \in \Omega_k} \nabla \ell(W_k^S, z_i) + \sqrt{2\eta\beta^{-1}} \mathcal{N}(0, I_d)$$

  - Very flexible assumptions
  - Bounds computable from the learning trajectory

- **Algorithmic stability**
  - Adapted to vanilla SGD (and also noisy variants)
  - Often restrictive and global assumptions (convex / Lipschitz / smooth / dissipativity)

The success of a method depends on the modelisation of the noise

Noisy variants and modeling of the noise

- For large batch SGD, the central limit theorem suggests to represent the gradient noise

$$\nabla \widehat{\mathcal{R}}_S(w) - \frac{1}{b} \sum_{i \in \Omega_k} \nabla \ell(w, z_i),$$

  as a Gausian[4] [5] with covariance $b^{-1}C(w)$
- Euler-Maruyama discretization of an SDE of the form:

$$\mathrm{d}W_t^S = -\nabla \widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sqrt{\frac{\eta}{b}} D(W_t^S)\mathrm{dB}_t, \quad DD^T = C.$$

---

[4] Jastrzebski et al., Three Factors Influencing Minima in SGD, 2017

[5] Mandt et al.,A Variational Analysis of Stochastic Gradient Algorithms, ICML 2017

Generalization bounds for Langevin SDEs

- A simpler and highly-studied model are the Langevin SDEs

$$\mathrm{d}W_t^S = -\lambda W_t^S \mathrm{d}t - \nabla \widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sqrt{2}\sigma \mathrm{dB}_t$$

- Generalization bounds for such algorithms are highy studied

Theorem 1 (Improved from Mou et al., 2017)

Assume that $\ell(w, z)$ is Σ^2-subgaussian and let $\tau := \lambda^{-1}$. Then, with probability at least $1 - \zeta$ over $S \sim \mu_z^{\otimes n}$,

$$\mathbb{E}\left[G_S(W_T^S)|S\right] \leq \frac{2\Sigma}{\sqrt{n}} \left\{ \frac{1}{\sigma^2} \int_0^T e^{-\frac{T-t}{\tau}} \mathbb{E}[\|\nabla \widehat{\mathcal{R}}_S(W_t^S)\|^2|S]\mathrm{d}t + \log \frac{3}{\zeta} \right\}^{\frac{1}{2}}$$

- Information-theoretic method
- Refined in many subsequent works [13, 9]

Sketch of proof: PAC-Bayesian bounds

- $\mathrm{d}W_t^S = -\lambda W_t^S \mathrm{d}t - \nabla \widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t.$

Sketch of proof: PAC-Bayesian bounds

- $\mathrm{d}W_t^S = -\lambda W_t^S \mathrm{d}t - \nabla \widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t$.
- $\rho_t^S := \mathrm{Law}(W_t^S)$ is the posterior distribution, u_t^S its density

Theorem 2 (Example of PAC-Bayesian bound (ours) [6])

Assume that ℓ is Σ^2-subgaussian. We have

$$\mathbb{P}_{S\sim\mu_z^{\otimes n}}\left(\forall t,\ \mathbb{E}_{W\sim\rho_t^S}\left[G_S(W)\right] \leq 2\Sigma\sqrt{\frac{\mathbf{KL}(\rho_t^S|\pi) + \log(3/\zeta)}{n}}\right) \geq 1-\zeta.$$

- π: a well-chosen prior distribution (data-independent), *e.g.*, the invariant distribution of a the Ornstein-Uhlenbeck process:

$$\mathrm{d}W_t = -\lambda W_t \mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t$$

- Reduces the problem to estimating a relative entropy
- One of the basic information-theoretic approaches

Sketch of proof: entropy flow for Langevin SDEs (2017)

- Key idea: compute the entropy flow

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{KL}(\rho_t^S|\pi) \leq -\frac{\sigma^2}{2}\mathscr{J}(\rho_t^S||\pi) + \frac{1}{2\sigma^2}\int \|\nabla\widehat{\mathcal{R}}_S\|^2 \mathrm{d}\rho_t^S.}$$

- $\mathscr{J}(\rho_t^S||\pi) := \int \|\nabla \log v_t\| \mathrm{d}\rho_t^S$ is the relative Fisher information

- By the **log-Sobolev inequality** for π:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{KL}(\rho_t^S|\pi) \leq -\lambda\mathbf{KL}(\rho_t^S|\pi) + \frac{1}{2\sigma^2}\int \|\nabla\widehat{\mathcal{R}}_S\|^2 \mathrm{d}\rho_t^S.$$

The result follows from Grönwall's lemma

Generalization bounds for SGLD (discrete-time version)

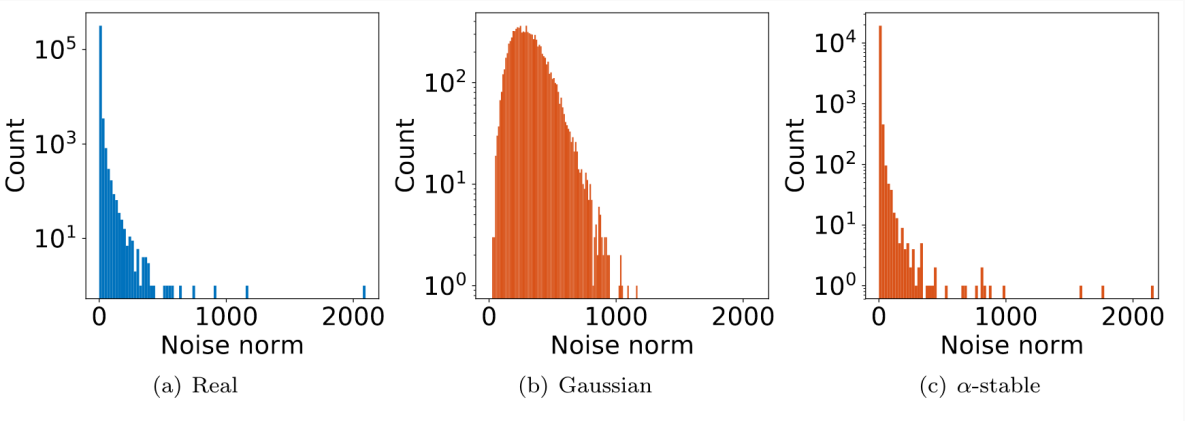
$$W_{k+1}^S = W_k^S - \eta_k \widehat{g}_k^S + \sqrt{2\eta_k \beta^{-1}} \mathcal{N}(0, I_d).$$

Paper	Method	Asmp. on $\ell(w,z)$	Bound ($\mathcal{O}$)
[14]	stability + IT	L. B.	$\frac{LB}{n}\sqrt{\beta \sum_k \eta_k}$
[14]	IT	L. SG. BS-1 R.	$\left\{\frac{\beta}{n}\sum_{k=1}^T e^{-R_k^T}\eta_k \mathbb{E}\left\|g_k\right\|^2\right\}^{\frac{1}{2}}$
[18]	IT	L. SG. BS-1	$L\sqrt{\frac{\beta}{n\zeta}\sum_{k=1}^T \eta_k}$
[15]	IT	SG. BS-b	$\left\{\frac{1}{bn}\sum_{k=1}^T \beta_k \eta_k \mathsf{Var}(g_k)\right\}^{\frac{1}{2}}$
[11]	IT	B. BS-n	$\frac{B}{n}\left\{\sum_{k=1}^T \eta_k \beta_k \left\|\zeta_k\right\|^2\right\}^{\frac{1}{2}}$
[16]	IT	SG. BS-1	$\left\{\frac{\beta}{n}\sum_{k=1}^T \eta_k \mathsf{Var}(g_k \vert w_k)\right\}^{\frac{1}{2}}$
[8]	stability	SG. D.	$\min\{1,\eta T\}\left(\sqrt{\eta} + \frac{1}{n\sqrt{\eta}}\right)$
[9]	IT	SG. D.	$\sqrt{\frac{\min(1,\eta T)}{n}}$

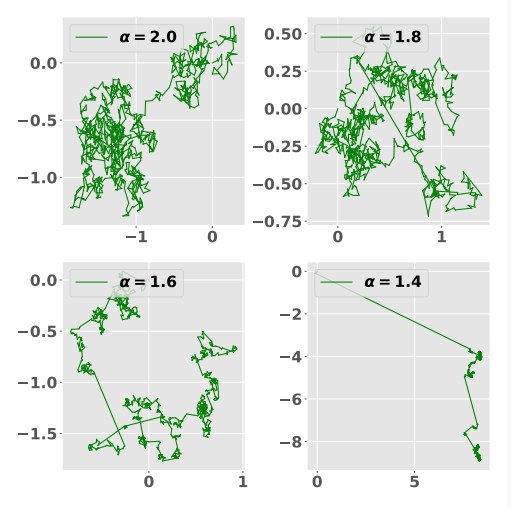
Table: L. Lipschitz, ∇-M. M-smooth, C. convex, B. bounded, SG. subgaussian, D. dissipatif, p.L. pseudo-Lipschitz, SC. strongly convex

Extension 1: Heavy-tailed noise

What does the gradient noise look like in practice? [20, 10]



(a) Real (b) Gaussian (c) α-stable



- Suggest to model the noise with heavy-tailed distributions[6] [7]
- Heavy-tailed SDE:

$$\mathrm{d}W_t^S = -\nabla\widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sigma_1\mathrm{d}\mathrm{L}_t^\alpha + \sqrt{2}\sigma_2\mathrm{d}B_t$$

- $\alpha \in (0, 2]$ is the tail index and L_t^α is a stable Lévy process

---

[6] Simsekli et al.,A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks, ICML 2019

[7] Gurguzbalaban et al., The Heavy-Tail Phenomenon in SGD, ICML 2021

Generalization bounds for heavy-tailed SDEs

- Complex dependence[8] between the generalization error and α

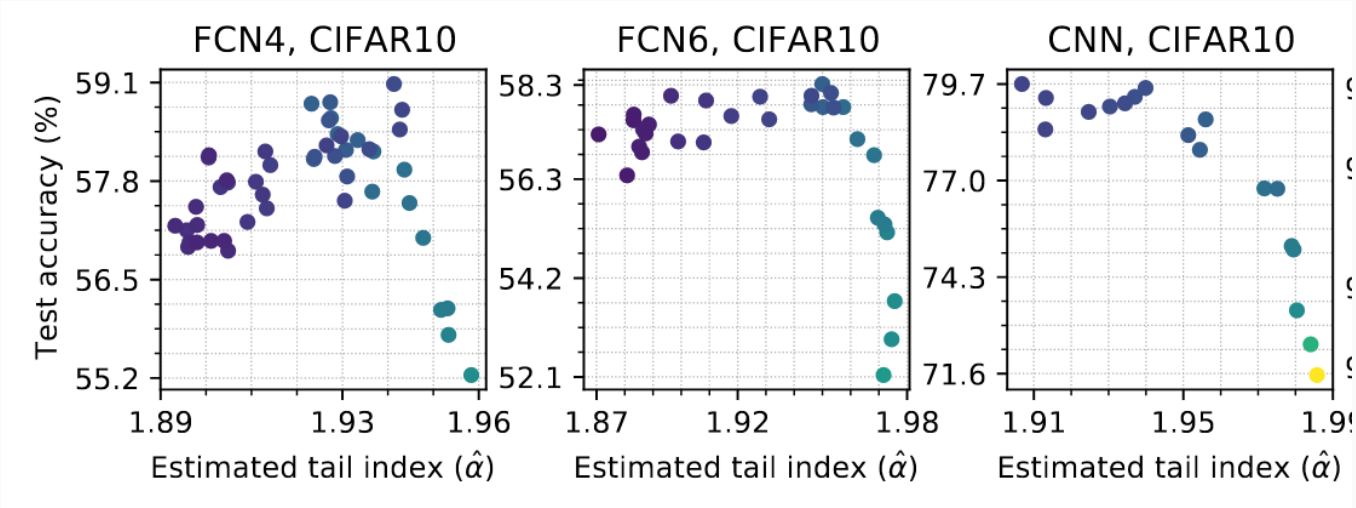


Figure: Test accuracy after convergence with respect to the estimated tail index

- Existing works [19, 21]:
  - Strong dependence on the ambient dimension
  - Complicated assumptions
  - Do not fully capture the relation between generalization and α

---

[8] Barsbey et al.,Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks, ICML 2019

Generalization bounds for heavy-tailed SDEs[9]

- Generalization bounds for $\sigma_2 > 0$ (easy) and $\sigma_2 = 0$ (hard)
- Under suitable assumptions, improved dimension dependence

Theorem 3 (informal)

Then, with probability at least $1-\zeta$ *over* $\mu_z^{\otimes n}$*, for all* $T > 0$*,*

$$\mathbb{E}\left[G_S(W_T^S)|S\right] \leq \frac{2\Sigma}{\sqrt{n}} \left\{ \frac{\mathrm{K}_{\alpha,d}}{\sigma_1^\alpha} \int_0^T \mathbb{E}[\|\nabla \widehat{\mathcal{R}}_S(W_t^S)\|^2|S]\mathrm{d}t + \log\frac{3}{\zeta} \right\}^{\frac{1}{2}},$$

with

$$\mathrm{K}_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right) d\Gamma\left(\frac{d}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right) R^{2-\alpha}} \underset{d\to\infty}{\sim} P_\alpha d^{1-\alpha/2}$$

where Γ *denotes the Euler Gamma function.* R *is a constant that we assume controls the regularity of some terms appearing in our theory.*

[9] **B. D.** and U. Şimşekli. Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation. ICML, 2024.[6]

Analysis

- Recovering the Gaussian case in the limit $\alpha \to 2^-$
- Prediction of a phrase transition of the tail index
- Heavy tails can be beneficial or harmful depending on the dimension

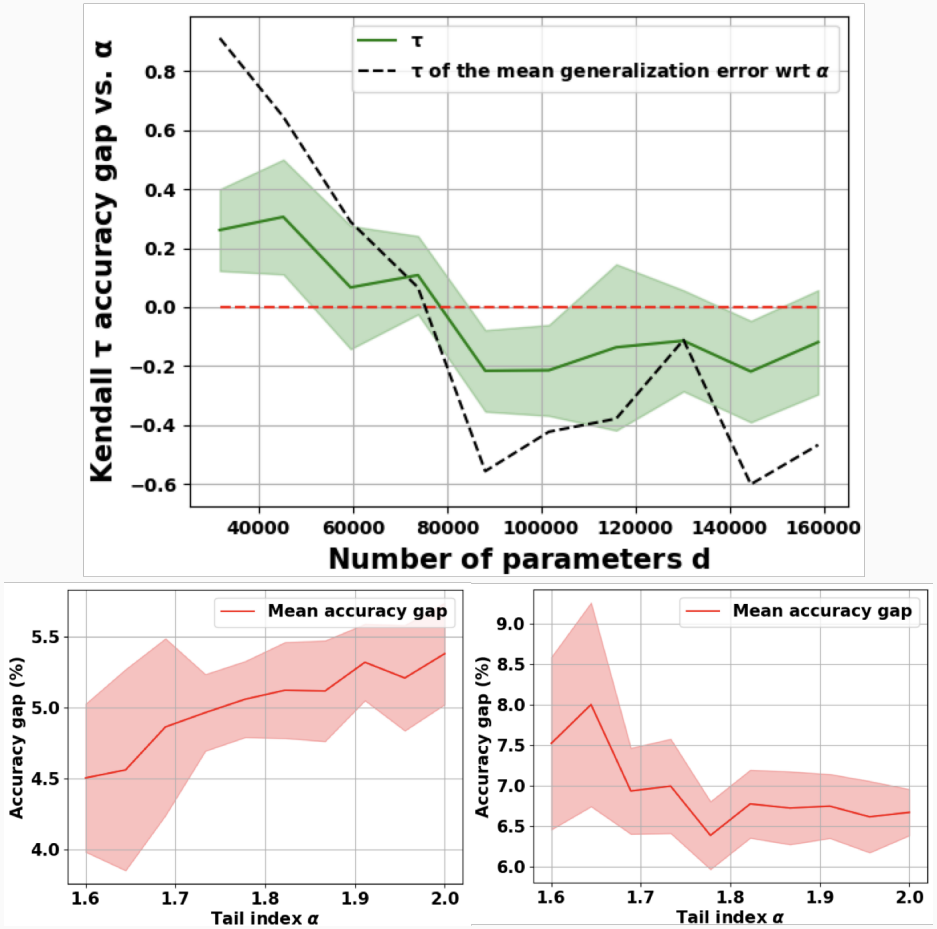


Figure: *(up)* Correlation (Kendall's τ) between α and the accuracy gap, for different values of d, MLP on MNIST.

Sketch of proof: PAC-Bayesian bounds again

- $\mathrm{d}W_t^S = -\lambda W_t^S \mathrm{d}t - \nabla \widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sigma_1 \mathrm{dL}_t^\alpha$.
- $\rho_t^S := \mathrm{Law}(W_t^S)$ is the posterior distribution, u_t^S its density

Theorem 4 (Example of PAC-Bayesian bound (ours) [6])

Assume that ℓ is Σ^2-subgaussian. We have

$$\mathbb{P}_{S\sim\mu_z^{\otimes n}}\left(\forall t,\ \mathbb{E}_{W\sim\rho_t^S}\left[G_S(W)\right] \leq 2\Sigma\sqrt{\frac{\mathbf{KL}(\rho_t^S|\pi) + \log(3/\zeta)}{n}}\right) \geq 1-\zeta.$$

- We take the prior distribution π as the invariant distribution of a the Lévy-Ornstein-Uhlenbeck process:

$$\mathrm{d}W_t = -\lambda W_t \mathrm{d}t + \sigma_1 \mathrm{dL}_t^\alpha$$

- Estimating the entropy flow is much more challenging than for Langevin SDEs

Sketch of proof: entropy flow for Lévy-driven SDEs

- We obtained a closed-form formula for the entropy flow

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{KL}(\rho_t^S|\pi) = -\sigma_1^\alpha \mathrm{B}_\alpha(v_t) - \int \nabla v_t \cdot \nabla \widehat{\mathcal{R}}_S \mathrm{d}\pi, \quad v_t := \frac{u_t^S}{\pi}.}$$

- Extension of the technique of Mou et al. (2017)
- Based on the fractional Fokker-Planck equations

$$\frac{\partial}{\partial t} u_t^S = -\sigma_1^\alpha (-\Delta)^{\frac{\alpha}{2}} u_t^S + \mathrm{div}(u_t^S \nabla \widehat{\mathcal{R}}_S),$$

- $\mathrm{B}_\alpha(v_t)$: New contraction term depending on α, the "Bregman integral"
- Our result follows from the analysis of this entropy flow

Extension of the entropy flow method

This technique:

- Requires few assumptions
- Provide explicit and empirically computable generalization bounds

Extension of the entropy flow method

This technique:

- Requires few assumptions
- Provide explicit and empirically computable generalization bounds

BUT:

- It is limited to Gaussian or Lévy noise structure
- It requires Fokker-Planck equations
- Does not apply to SGD and most of its variants

Can we extend the entropy flow method to a more general class of algorithms?

Extension of the entropy flow method[10]

Consider a Markov algorithm (such as SGD): $X_{k+1}^S := F(X_k^S, U_k, S)$

- S is the data, U_k is the noise

[10] **B. D.**, M. Haddouche, G. Deligiannidis, U. Şimşekli, Generalization Bounds for Markov Algorithms through Entropy Flow Computations 2025 [5]

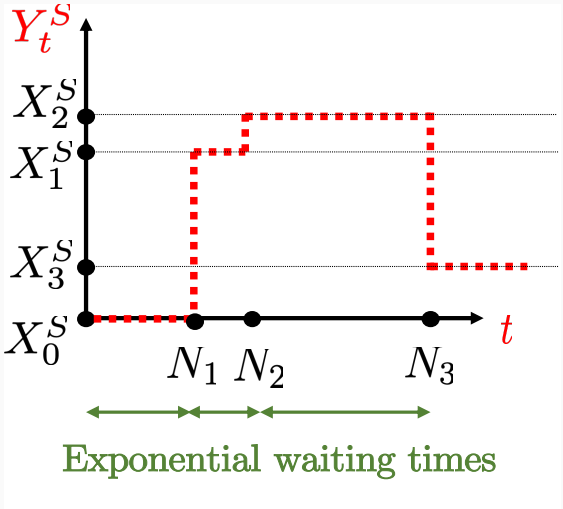
Extension of the entropy flow method[10]

Consider a Markov algorithm (such as SGD): $X_{k+1}^S := F(X_k^S, U_k, S)$

- S is the data, U_k is the noise
- We approximate it with a continuous-time process via Poissonization:

$$\boxed{Y_t^S := X_{N_t}^S,}$$

- with N_t a Poisson process



- The prior π: invariant distribution of a well-chosen Markov kernel P

[10] **B. D.**, M. Haddouche, G. Deligiannidis, U. Şimşekli, Generalization Bounds for Markov Algorithms through Entropy Flow Computations 2025 [5]

Extension of the entropy flow method

Theorem 5 (Entropy flow formula)

Let $\rho_t^S := \mathrm{Law}(Y_t^S)$ *and* $v_t := \mathrm{d}\rho_t^S/\mathrm{d}\pi$, *with* π *the prior distribution*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{KL}(\rho_t^S|\pi) = \underbrace{\int (P_S - P)(\log v_t)v_t \mathrm{d}\pi}_{\text{Expansion term } \Delta_{P,P_S}(v_t)} - \underbrace{\int v_t(I-P)(\log v_t)\mathrm{d}\pi}_{\text{Contraction term } \mathcal{E}_{\pi,P}(\log v_t, v_t) \geq 0}$$

Extension of the entropy flow method

Theorem 5 (Entropy flow formula)

Let $\rho_t^S := \mathrm{Law}(Y_t^S)$ *and* $v_t := \mathrm{d}\rho_t^S/\mathrm{d}\pi$, *with* π *the prior distribution*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{KL}(\rho_t^S|\pi) = \underbrace{\int (P_S - P)(\log v_t)v_t \mathrm{d}\pi}_{\text{Expansion term } \Delta_{P,P_S}(v_t)} - \underbrace{\int v_t(I-P)(\log v_t)\mathrm{d}\pi}_{\text{Contraction term } \mathcal{E}_{\pi,P}(\log v_t, v_t) \geq 0}$$

Expansion term: $\Delta_{P,P_S}(v_t)$

- Contribution of a single iteration to the entropy flow
- Measures the difference between P an P_S

Contraction term: $\mathcal{E}_{\pi,P}(\log v_t, v_t)$

- Dirichlet form associated with the prior kernel P
- Modified log-Sobolev inequalities [2], for suitable priors

$$\mathcal{E}_{\pi,P}(\log v_t, v_t) \geq \gamma \mathbf{KL}(\rho_t^S|\pi)$$

Extension of the entropy flow method

Theorem 6 (informal)

Assume $\ell(w,\cdot)$ is Σ^2-subgaussian, with probability at least $1-\zeta$, for all T simultaneously,

$$\mathbb{E}\left[G_S(Y_t^S)|S\right] \leq \frac{2\Sigma}{\sqrt{n}}\left\{\int_0^T e^{-\gamma(T-t)}\Delta_{P,P_S}(v_t)\mathrm{d}t + \log\frac{3}{\zeta}\right\}^{\frac{1}{2}}.$$

- We analyze the expansion term for noisy algorithms

$$\Delta_{P,P_S}(v_t) \leq \int \inf_{q>0}\left(\frac{1}{q}\mathbf{KL}(\delta_x P_S|\delta_x P) + \mathcal{O}(q)\right)\mathrm{d}\rho_t^S(x),$$

- We also obtain results for non-noisy algorithms

Applications:

- Recovered Mou et al. (2017) for SGLD

$$X_{k+1}^S = (1 - \lambda\eta)X_k^S - \eta\widehat{g}_S(X_k^S, U_k) + \sigma\xi_k, \quad \xi_k \sim \mathcal{N}(0, I_d),$$

- Improved bounds for SGD with perturbed last iterate [16]

$$X_{k+1}^S = (1 - \lambda\eta)X_k^S - \eta\widehat{g}_S(X_k^S, U_k), \quad \overline{X}_K^S := X_k^S + \sigma\mathcal{N}(0, I_d).$$

- First bounds for SGD with noise injection

$$X_{k+1}^S = X_k^S - \eta\nabla\widehat{\mathcal{R}}_S(X_k^S + \sigma\xi_k), \quad (\xi_k)_{k\in\mathbb{N}} \sim \mathcal{N}(0, I_d)^{\otimes n}$$

- New bounds for vanilla SGD under specific regularity conditions

Conclusion & Future works

Our other related works:

- Uniform generalization bounds for Langevin SDEs / SGLD[11]
- Differential privacy of heavy-tailed SDEs[12]
- Analysis of the score matching gap for diffusion models [13]

Future works:

- Extension to more complex algorithms
- Use more realistic covariance structures

Thank you!

---

[11] B.D., Paul Viallard, Umut Şimşekli, George Deligiannidis, Uniform Generalization Bounds over Data-Dependent Hypothesis sets via a PAC-Bayesian Theory on Random Sets, 2024 [7]

[12] B.D., Umut Şimşekli, Jian Wang, Sinan Yildirim, Lingjiong Zhu, Rényi Differential Privacy for Heavy-Tailed SDEs via Fractional Poincaré Inequalities 2025.[4]

[13] B.D., Tyler Farghly, Maxime Haddouche, Alain Durmus, Umut Şimşekli, Tightening the Score Matching Gap for Diffusion models ICML 2026 [3]

References I

[1] P. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 2002.

[2] P. Diaconis and L. Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

[3] B. Dupuis, T. Farghly, M. Haddouche, A. O. Durmus, and U. Simsekli. Tightening the score matching gap for diffusion models. 2026.

[4] B. Dupuis, M. Gürbüzbalaban, U. Simsekli, J. Wang, S. Yildirim, and L. Zhu. Rényi differential privacy for heavy-tailed SDEs via fractional Poincaré inequalities, 2025.

[5] B. Dupuis, M. Haddouche, G. Deligiannidis, and U. Simsekli. Generalization bounds for markov algorithms through entropy flow computations, 2026.

[6] B. Dupuis and U. Şimşekli. Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation. In *International Conference on Machine Learning (ICML)*, 2024.

References II

[7] B. Dupuis, P. Viallard, G. Deligiannidis, and U. Simsekli. Uniform Generalization Bounds on Data-Dependent Hypothesis Sets via PAC-Bayesian Theory on Random Sets. *Journal of Machine Learning Research*, 25(409):1–55, 2024.

[8] T. Farghly and P. Rebeschini. Time-independent Generalization Bounds for SGLD in Non-convex Settings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[9] F. Futami and M. Fujisawa. Time-Independent Information-Theoretic Generalization Bounds for SGLD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[10] M. Gürbüzbalaban, U. Şimşekli, and L. Zhu. The Heavy-Tail Phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, 2021.

[11] M. Haghifam, J. Negrea, A. Khisti, D. Roy, and G. K. Dziugaite. Sharpened Generalization Bounds Based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three Factors Influencing Minima in SGD. *arXiv preprint arXiv:1711.04623*, 2018.

References III

[13] J. Li, X. Luo, and M. Qiao. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[14] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In *Conference On Learning Theory (COLT)*, 2018.

[15] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. Roy. Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[16] G. Neu, G. K. Dziugaite, M. Haghifam, and D. Roy. Information-Theoretic Generalization Bounds for Stochastic Gradient Descent. In *Conference on Learning Theory (COLT)*, 2021.

[17] A. Orvieto, H. Kersting, F. Proske, F. Bach, and A. Lucchi. Anticorrelated noise injection for improved generalization. In *International Conference on Machine Learning*, volume 39, 2022.

[18] A. Pensia, V. Jog, and P.-L. Loh. Generalization Error Bounds for Noisy, Iterative Algorithms. *IEEE International Symposium on Information Theory (ISIT)*, 2018.

References IV

[19] A. Raj, L. Zhu, M. Gurbuzbalaban, and U. Simsekli. Algorithmic stability of heavy-tailed sgd with general loss functions. In *International Conference on Machine Learning*, pages 28578–28597, 2023.

[20] U. Şimşekli, L. Sagun, and M. Gürbüzbalaban. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019.

[21] U. Şimşekli, O. Sener, G. Deligiannidis, and M. Erdogdu. Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Heavy-tailed entropy flow (details)

$$\mathrm{d}W_t^S = -\lambda W_t^S \mathrm{d}t - \nabla \widehat{\mathcal{R}}_S(W_t^S)\mathrm{d}t + \sigma_1 \mathrm{dL}_t^\alpha + \sqrt{2}\sigma_2 \mathrm{d}B_t.$$

Under appropriate regularity assumptions, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{KL}(\rho_t^S|\pi) = -\sigma_2^2 \mathscr{J}(\rho_t^S||\pi) - \sigma_1^\alpha \mathrm{B}_\Phi^\alpha(v_t) - \int \nabla v_t \cdot \nabla \widehat{\mathcal{R}}_S \mathrm{d}\pi,$$

where $\mathrm{B}_\Phi^\alpha(v)$, with $\Phi(x) := x\log(x)$ the "Bregman integral", which is defined as

$$\boxed{\mathrm{B}_\Phi^\alpha(v) := C_{\alpha,d} \int\int \mathrm{D}_\Phi(v(x), v(x+z))\mathrm{d}\nu_\alpha(z)\mathrm{d}\pi(x), \quad \mathrm{d}\nu_\alpha(z) := \frac{\mathrm{d}z}{\|z\|^{d+\alpha}},}$$

and the constant $C_{\alpha,d}$ is given by

$$C_{\alpha,d} := \alpha 2^{\alpha-1}\pi^{-d/2}\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(1-\frac{\alpha}{2}\right)}.$$

Multifractal case: $\sigma_2 > 0$

Theorem 7

Let $\Lambda := \mathbf{KL}(\rho_0|\pi)$ and assume that the loss is Σ^2-subgaussian. With probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, we have

$$G_S(T) \leq \frac{2\Sigma}{\sqrt{n}} \left\{ \frac{1}{4\sigma_2^2} \int_0^T \mathbb{E}\left[\|\nabla \widehat{\mathcal{R}}_S(W_t^S)\|^2 | S \right] \mathrm{d}t + \log(3/\zeta) + \Lambda \right\}^{1/2}.$$

Proof: Combine the PAC-Bayes bound, the entropy flow formula, and Young's inequality.

Analysis of the Bregman integral

We show that there exists a function, $\mathrm{J}_{\Phi,v} : [0, +\infty) \to [0, +\infty)$, such that $\mathrm{J}_{\Phi,v}$ is non-negative, continuous, satisfies $\mathrm{J}_{\Phi,v_t}(0) = \mathscr{J}(\rho_t^S || \pi)$ and we have

$$\mathrm{B}_\Phi^\alpha(v) = C_{\alpha,d} \frac{\sigma_{d-1}}{2d} \int_0^\infty \mathrm{J}_{\Phi,v}(r) \frac{dr}{r^{\alpha-1}}, \quad \sigma_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)},$$

- Related to weighted Bourgain-Brezis-Mironescu types formulas
- We assume that we can control the modulus of continuity of this function:

Assumption 1.1

There exists $R > 0$ such that, for all $t > 0$ and $\mu_z^{\otimes n}$-almost all $S \in \mathcal{Z}^n$,

$$\forall r \in [0, R],\ 2\mathrm{J}_{\Phi,v_t}(r) \geq \mathrm{J}_{\Phi,v_t}(0).$$

Extension of the entropy flow method

Theorem 8 (informal)

Assume $\ell(w,\cdot)$ is Σ^2-subgaussian, with probability at least $1-\zeta$, for all T simultaneously,

$$\mathbb{E}\left[G_S(Y_t^S)|S\right] \leq \frac{2\Sigma}{\sqrt{n}}\left\{\int_0^T e^{-\gamma(T-t)}\Delta_{P,P_S}(v_t)\mathrm{d}t + \log\frac{3}{\zeta}\right\}^{\frac{1}{2}}.$$

- We analyze the expansion term for noisy algorithms

$$\Delta_{P,P_S}(v_t) \leq \int \inf_{q>0}\left(\frac{1}{q}\mathbf{KL}(\delta_x P_S|\delta_x P) + \mathcal{O}(q)\right)\mathrm{d}\rho_t^S(x),$$

- We also obtain results for non-noisy algorithms

Application: SGLD

Stochastic Gradient Langevin Dynamics (SGLD):

$$X_{k+1}^S = (1 - \lambda\eta)X_k^S - \eta\widehat{g}_S(X_k^S, U_k) + \sigma\xi_k, \quad \xi_k \sim \mathcal{N}(0, I_d),$$

where $\widehat{g}_S(X_k^S, U_k)$ is a stochastic gradient.

Theorem 9

Suppose that $\eta\lambda < 1$. Assume that the loss is Σ^2-subgaussian. Then, with probability at least $1 - \zeta$ over $S \sim \mu_z^{\otimes n}$, for all $T > 0$,

$$\mathbb{E}\left[G_S(Y_t^S)|S\right] \leq \frac{2\Sigma}{\sqrt{n}}\left(\frac{\eta^2}{\sigma^2}\int_0^T e^{-\lambda\eta(T-t)}\mathbb{E}[\|\widehat{g}_S(Y_t^S)\|^2|S]\mathrm{d}t + \log\frac{3}{\zeta}\right)^{1/2}$$

- Recovers a Poissonized version of Mou et al. (2017) [14]
- The exponential decay comes from the ridge regularization
- Prior process

$$X_{k+1} = (1 - \lambda\eta)X_k + \sigma\xi_k$$

Application: SGD with perturbed last iterate

- Vanilla SGD algorithm $X^S_{k+1} = (1-\lambda\eta)X^S_k - \eta\widehat{g}_S(X^S_k, U_k)$
- Adding noise to the last iterate, making it more suitable for information-theoretic approaches [16]

$$\bar{Y}^S_t := \bar{X}^S_{N_t} = Y^S_t + \sigma\xi_{N_t},$$

Theorem 10

Take $\pi := \mathcal{N}(0, \sigma^2 I_d)$. Assume that the loss $\ell(w,z)$ is Σ^2-subgaussian wrt $z \sim \mu_z$, and that the initial distribution μ_0 has its support included in the ball $B(0, L/\lambda)$. Let $\tau = 1/(\lambda\eta)$ with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, for all $T > 0$,

$$\mathbb{E}\left[G_S(\bar{Y}^S_T)|S\right] \leq \frac{2\Sigma}{\sqrt{n}}\left\{\int_0^T \frac{\eta L e^{-\frac{T-t}{\tau}}}{\lambda\sigma^2}\sqrt{\mathbb{E}[\|\widehat{g}_S(Y^S_t, U)\|^2\big|S]}\mathrm{d}t + \log\frac{3}{\zeta}\right\}^{\frac{1}{2}}$$

with $U \sim \mu_U$ independent of Y^S_t.

Application: SGD with noise injection

Proposed to regularize the dynamics towards flat minima [17]:

$$X_{k+1}^S = X_k^S - \eta \nabla \widehat{\mathcal{R}}_S(X_k^S + \sigma \xi_k), \quad (\xi_k)_{k \in \mathbb{N}} \sim \mathcal{N}(0, I_d)^{\otimes n},$$

Theorem 11

Assume that $\ell(\cdot, z) \in \mathcal{C}^2(\mathbb{R}^d)$ and is γ-strongly convex (with $\gamma\eta < 1$) and gradient-Lipschitz for all $z \in \mathcal{Z}$, and that $\ell(w, z)$ is Σ^2-subgaussian with respect to $z \sim \mu_z$. Let $\tau := 1/(\eta\gamma)$, with probability at least $1 - \zeta$,

$$\mathbb{E}\left[G_S(Y_t^S)|S\right] \leq \frac{4\Sigma}{\sqrt{n}} \left\{ \int_0^T e^{-\frac{T-t}{\tau}} \mathbb{E}_{\bar{\rho}_t^S}[\Delta \widehat{\mathcal{L}}_S + \frac{1}{\sigma^2} \|\nabla \widehat{\mathcal{L}}_S\|^2] \mathrm{d}t + \log \frac{3}{\zeta} \right\}^{\frac{1}{2}}$$

where $\widehat{\mathcal{L}}_S(y) = \gamma^{-1} \widehat{\mathcal{R}}_S(y) - \|y\|^2 / 2$ and $\pi = \mathcal{N}(0, \sigma^2 \gamma\eta/(2 - \gamma\eta) I_d)$.

Application: vanilla SGD

$$X_{k+1}^S = (1 - \lambda\eta)X_k^S - \eta\widehat{g}_S(X_k^S, U_k)$$

Theorem 12

Assume that $\ell(w, z)$ is Σ^2-subgaussian and that there exist $\sigma, a > 0$ such that for all $t \in [0, T]$, $\left\|\nabla \log u_t^S(x) - \sigma x\right\| \leq a \|x\| + b$. Let $\tau := 1/(\lambda\eta)$, with probability at least $1 - \zeta$ over $S \sim \mu_z^{\otimes n}$, that for all $T > 0$,

$$\mathbb{E}\left[G_S(Y_T^S)|S\right] \lesssim \frac{2\Sigma}{\sqrt{n}} \left\{ \int_0^T e^{-\frac{T-t}{\tau}} \mathrm{N}_S(t) \left(\mathrm{N}_S(t) + \mathrm{M}_S(t)+\right) \mathrm{d}t + \log \frac{3}{\zeta} \right\}^{\frac{1}{2}}$$

with $\mathrm{N}_S(t)^2 := \eta^2 \mathbb{E}[\|\widehat{g}_S(Y_t^S, U)\|^2|S] + 2\sigma^2\lambda\eta d$ and $\mathrm{M}_S(t)^2 := \mathbb{E}[\|Y_t^S\|^2|S]$.