Temperature is All You Need

for Generalization in Langevin Dynamics and other **Data Dependent Markov Processes**

Nati Srebro
TTIC


Itamar Harel


Yonatan Wolanowsky


Daniel Soudry

Technion


Gal Vardi

Weizman

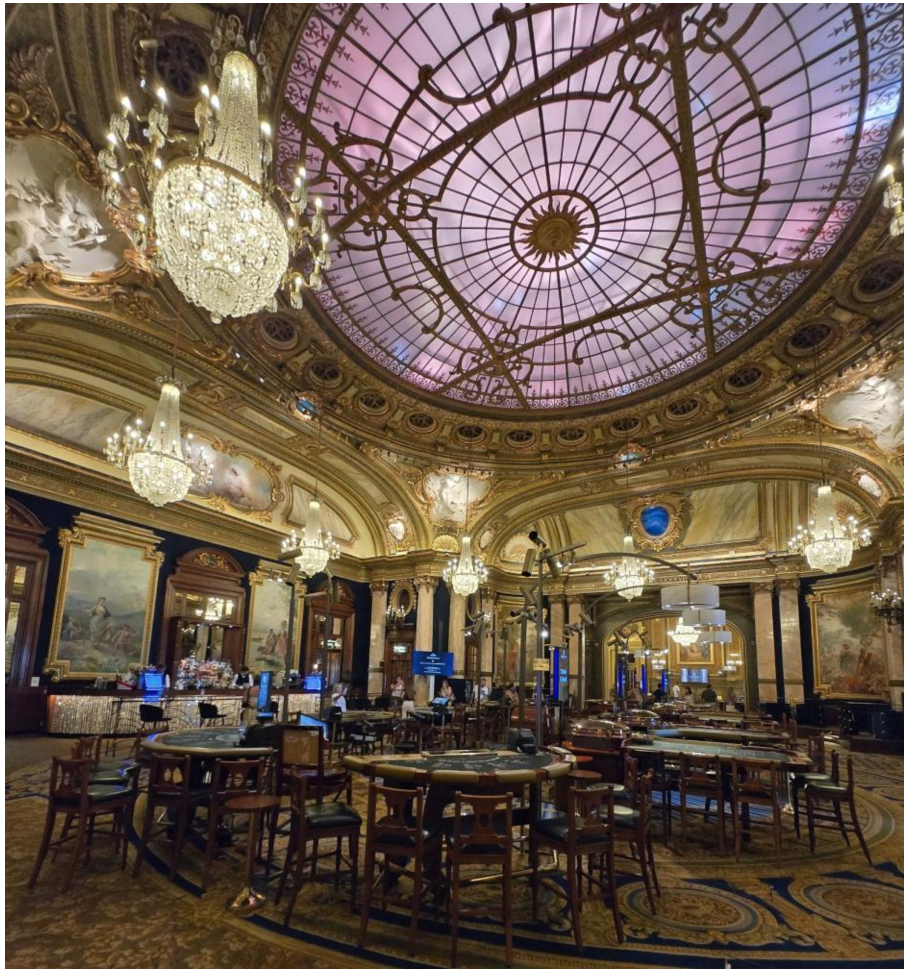
The journey is half the reward

- Collect Data $S = \{z_1, z_2, \dots, z_m\} \sim \mathcal{D}^m$
- Learn h by running a time-invariant Markov chain based on S:
  - Init $h \sim p_0(\cdot; S)$
  - $h_{t+1} | h_t \sim r(\cdot \mid \cdot; S)$ (or in continuous time)

Examples:

SGD: $h_{t+1} = h_t - \nabla loss(z; h_t), \;\; z \sim S$

Langevin: $\mathrm{d}h_t = -\nabla L_S(h_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t$



- Collect Data $S = \{z_1, z_2, \dots, z_m\} \sim \mathcal{D}^m$
- Learn h by running a time-invariant Markov chain based on S:
  - Init $h \sim p_0(\cdot; S)$
  - $h_{t+1} | h_t \sim r(\cdot \mid \cdot; S)$ (or in continuous time)

Can we bound the generalization gap?

$$|L_S(h_t) - L_\mathcal{D}(h_t)|$$

$L_S(h) = \frac{1}{m} \sum loss(h; z_i)$ $\qquad$ $L_\mathcal{D}(h) = \mathbb{E}_{z \sim \mathcal{D}} loss(h; z)$

Examples:

SGD: $h_{t+1} = h_t - \nabla loss(z; h_t), \; z \sim S$

Langevin: $\mathrm{d}h_t = -\nabla L_S(h_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t$

Main tool: PAC-Bayes Bound

For any base measure/"prior" ν (over h),
with probability $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$,
for any $p(h; S)$,

$$\mathbb{E}_{h \sim p(\cdot; S)}[L_\mathcal{D}(h) - L_S(h)] \leq \sqrt{\frac{KL(p(\cdot; S) \| \nu) + \ln {}^1/_\delta}{2m}}$$

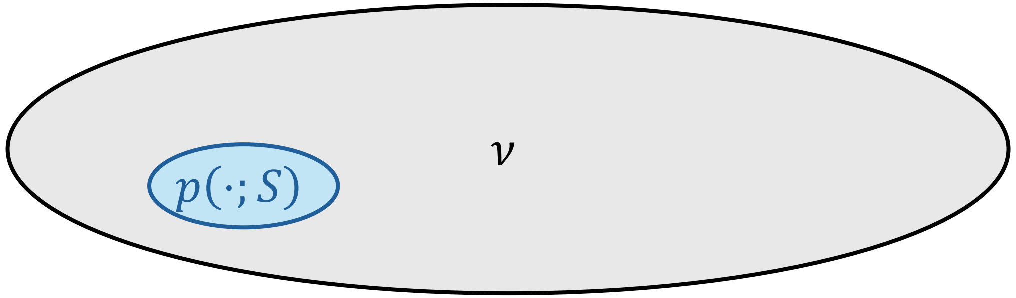
for $0 \leq loss \leq 1$

Special case: $p(h; S) = \delta_{h_S}$, point mass on single h_S

➔ $KL(p \| \nu) = -\ln \nu(h_S) \propto$ #bits to describe h_S

$= \ln |\mathcal{H}|$

If $\nu(h)$ uniform on $\mathcal{H}$



- Collect Data $S = \{z_1, z_2, \dots, z_m\} \sim \mathcal{D}^m$
- Learn h by running a time-invariant Markov chain based on S:
  - Init $h \sim p_0(\cdot; S)$
  - $h_{t+1} | h_t \sim r(\cdot \mid \cdot; S)$ (or in continuous time)

Can we bound the generalization gap?

$$|L_S(h_t) - L_{\mathcal{D}}(h_t)|$$

$L_S(h) = \frac{1}{m} \sum loss(h; z_i)$ $\qquad$ $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}} loss(h; z)$

Examples:

SGD: $h_{t+1} = h_t - \nabla loss(z; h_t), \;\; z \sim S$

Langevin: $\mathrm{d}h_t = -\nabla L_S(h_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t$

Main tool: PAC-Bayes Bound

For any base measure/"prior" ν (over h),
with probability $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$,
for any $p(h; S)$,

$$\mathbb{E}_{h \sim p(\cdot; S)}[L_{\mathcal{D}}(h) - L_S(h)] \leq \sqrt{\frac{KL(p(\cdot; S) \| \nu) + \ln {}^1/_\delta}{2m}}$$

for $0 \leq loss \leq 1$

At $t = \infty$, perhaps $\mathrm{d}p_\infty \propto e^{-\Psi(h)}\mathrm{d}\nu$
E.g. for (regularized) Langevin $\Psi(h) = \beta L_S(h)$

$$\Rightarrow KL(p_\infty \| \nu) + \overbrace{KL(\nu \| p_\infty)}^{\geq 0} = \beta \mathbb{E}_\nu L_S(h) - \overbrace{\beta \mathbb{E}_{p_\infty} L_S(h)}^{\geq 0}$$

$$\mathbb{E}[\boldsymbol{L_{\mathcal{D}}(h_\infty)} - \boldsymbol{L_S(h_\infty)}] \leq \sqrt{\frac{\boldsymbol{\beta}\, \mathbb{E}_\nu L_S(h) \; + \; ln\, {}^1/_\delta}{\boldsymbol{2m}}}$$

Theorem: if p is **Gibbs** wrt q, i.e. $\mathrm{d}p \propto e^{-\Psi}\mathrm{d}q$ then

$$KL(p \| q) + KL(q \| p) = \mathbb{E}_q \Psi - \mathbb{E}_p \Psi$$

Proof:

$$\cdots = \mathbb{E}_p \ln \frac{\mathrm{d}p}{\mathrm{d}q} + \mathbb{E}_q \ln \frac{\mathrm{d}p}{\mathrm{d}q} = \mathbb{E}_p[-\Psi - \cancel{\ln Z}] + \mathbb{E}_q[\Psi + \cancel{\ln Z}]$$

$\mathrm{d}p = \frac{1}{Z} e^{-\Psi} \mathrm{d}q$

Second Law of Thermodynamics a la Thomas Cover:

for any stationary p_∞: $\quad \boldsymbol{KL(p_{t+1}\|p_\infty) \leq KL(p_t\|p_\infty)}$

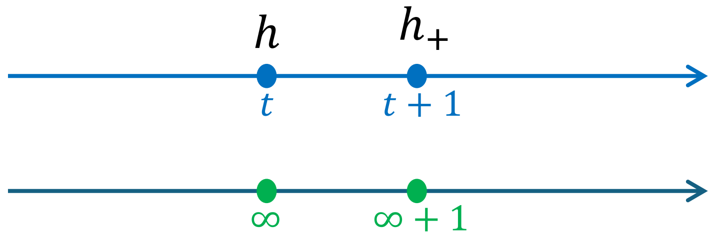
in any time-invariant Markov Chain

[**Which processes satisfy the second law?**, in *Physical Origins of Time Asymmetry* 1994]

<u>Proof</u>: Consider two joint distributions over pairs of variables in the chain

$p(h, h_+)$ where $h \sim p_t$ and $h_+ | h \sim r(\cdot \mid \cdot)$

$q(h, h_+)$ where $h \sim p_\infty$ and $h_+ | h \sim r(\cdot \mid \cdot)$



data processing

$$KL(p_{t+1}\|p_\infty) = KL\big(p(h_+)\|q(h_+)\big) \leq KL\big(p(h, h_+)\|q(h, h_+)\big)$$

$$= KL\big(p(h)\|q(h)\big) + \mathbb{E}_{h\sim p}\cancel{KL\big(r(h_+|h)\|r(h_+|h)\big)} = KL(p_t\|p_\infty)$$

Second Law of Thermodynamics a la Thomas Cover:

for any stationary p_∞: $\boldsymbol{KL(p_t \| p_\infty) \leq KL(p_0 \| p_\infty)}$

$$KL_\mu(p\|q) = \mathbb{E}_\mu\left[\ln \frac{\mathrm{d}p}{\mathrm{d}q}\right]$$

$$\boldsymbol{KL(p_t\|\nu)} = KL(p_t\|p_\infty) + KL_{p_t}(p_\infty\|\nu) \leq KL(p_0\|p_\infty) + KL_{p_t}(p_\infty\|\nu)$$
$$= KL(p_0\|\nu) + KL_{p_0}(\nu\|p_\infty) + KL_{p_t}(p_\infty\|\nu) \leq \boldsymbol{KL(p_0\|\nu)} + \mathbb{E}_{\boldsymbol{p_0}}\boldsymbol{\Psi} \underbrace{-\mathbb{E}_{p_t}\Psi}_{\geq 0}$$

$\mathrm{d}p_\infty \propto e^{-\Psi}\mathrm{d}\nu$

Theorem: if p is **Gibbs** wrt q, i.e. $\mathrm{d}p \propto e^{-\Psi}\mathrm{d}q$ then

$$KL_\mu(p\|q) + KL_\eta(q\|p) = \mathbb{E}_\eta\Psi - \mathbb{E}_\mu\Psi$$

Proof:

$$\cdots = \mathbb{E}_\mu \ln\frac{\mathrm{d}p}{\mathrm{d}q} + \mathbb{E}_\eta \ln\frac{\mathrm{d}p}{\mathrm{d}q} = \mathbb{E}_\mu[-\Psi - \ln Z] + \mathbb{E}_\eta[\Psi + \ln Z]$$

If exists stationary dist $\mathrm{d}p_\infty \propto e^{-\Psi}\mathrm{d}\nu$, $\Psi \geq 0$

➔ $KL(p_t\|\nu) \leq KL(p_0\|\nu) + \mathbb{E}_{p_0}\Psi$

Conclusion: For any time-inv data-depndnt Markov Process with some stationary distribution $p_\infty(\cdot; S)$ that is Gibbs w.r.t. a fixed (non data dependent) ν with potential $\Psi(h;S) \geq 0$, with prob $\geq 1-\delta$ over $S \sim \mathcal{D}^m$:

$$\mathbb{E}_{h_t}[L_\mathcal{D}(h_t) - L_S(h_t)] \leq \sqrt{\frac{\mathbb{E}_{h\sim p_0}\Psi(h) + KL(p_0\|\nu) + \ln {}^1\!/_\delta}{m}} \leq \sqrt{\frac{\beta + \ln {}^1\!/_\delta}{m}}$$

for $0 \leq loss \leq 1$

If p_0 not data dependent,
$\mathrm{d}p_\infty \propto e^{-\beta\mathcal{L}_S}\mathrm{d}p_0$, $\mathbb{E}_{p_0}\mathcal{L}_s \leq 1$

In-Expectation PAC-Bayes Bound

For any (data independent) ν and data dependent p_S, with probability $\geq 1-\delta$ over $S \sim \mathcal{D}^m$

$$\mathbb{E}_{h\sim p_S}[L_{\mathcal{D}}(h) - L_S(h)] \leq \sqrt{\frac{KL(p_S\|\nu) + \ln {}^1\!/_\delta}{m}}$$

for $0 \leq loss \leq 1$

$$D_\infty(p\|q) = \sup_p \ln \frac{\mathrm{d}p}{\mathrm{d}q}$$

Single Sample PAC-Bayes Bound

For any (data independent) ν and data dependent p_S,

with probability $\geq 1-\delta$ over $S \sim \mathcal{D}^m$ **AND** $h\sim p_S$

$$L_{\mathcal{D}}(h) - L_S(h) \leq \sqrt{\frac{\ln \frac{\mathrm{d}p_S}{\mathrm{d}\nu}(h) + \ln {}^1\!/_\delta}{2m}} \leq \sqrt{\frac{D_\infty(p_S\|\nu) + \ln {}^1\!/_\delta}{2m}}$$

based on [Alquier 2024]

$$kl(L_S(h)\|L_{\mathcal{D}}) \leq \frac{D_\infty(p_S\|\nu) + \ln {}^{2m}\!/_\delta}{m}$$

for $0 \leq loss \leq 1$

Second Law of Thermodynamics (pointwise):

$$D_\infty(p_t \| p_\infty) \le D_\infty(p_0 \| p_\infty)$$

Theorem: if p is **Gibbs** wrt q, i.e. $\mathrm{d}p \propto e^{-\Psi}\mathrm{d}q$ then

$$D_\infty(p\|q) + D_\infty(q\|p) = \sup_q \Psi - \inf_p \Psi$$

If exists stationary dist $\mathrm{d}p_\infty \propto e^{-\Psi}\mathrm{d}\nu,\ \Psi \ge 0$

➔ $D_\infty(p_t\|\nu) \le D_\infty(p_0\|\nu) + \sup \Psi$

$$D_\infty(p\|q) = \sup_p \ln \frac{\mathrm{d}p}{\mathrm{d}q}$$

Single Sample PAC-Bayes Bound

For any (data independent) ν and data dependent p_S,

with probability $\ge 1-\delta$ over $S \sim \mathcal{D}^m$ **AND** $h \sim p_S$

$$L_{\mathcal{D}}(h) - L_S(h) \le \sqrt{\frac{\ln \frac{\mathrm{d}p_S}{\mathrm{d}\nu}(h) + \ln {}^1\!/_\delta}{2m}} \le \sqrt{\frac{D_\infty(p_S\|\nu) + \ln {}^1\!/_\delta}{2m}}$$

$$kl(L_S(h)\|L_{\mathcal{D}}) \le \frac{D_\infty(p_S\|\nu) + \ln {}^{2m}\!/_\delta}{m}$$

for $0 \le loss \le 1$

Second Law of Thermodynamics (pointwise):

$$D_\infty(p_t \| p_\infty) \le D_\infty(p_0 \| p_\infty)$$

Theorem: if p is **Gibbs** wrt q, i.e. $\mathrm{d}p \propto e^{-\Psi}\mathrm{d}q$ then

$$D_\infty(p\|q) + D_\infty(q\|p) = \sup_q \Psi - \inf_p \Psi$$

If exists stationary dist $\mathrm{d}p_\infty \propto e^{-\Psi}\mathrm{d}\nu,\ \Psi \ge 0$

➔ $D_\infty(p_t\|\nu) \le D_\infty(p_0\|\nu) + \sup \Psi$

$$D_\infty(p\|q) = \sup_p \ln \frac{\mathrm{d}p}{\mathrm{d}q}$$

Conclusion: For any time-inv data-depndnt Markov Process with some stationary distribution $p_\infty(\cdot;S)$ that is Gibbs w.r.t. a fixed (non data depndnt) ν with potential $0 \le \Psi(h;S) \le \beta$, **with prob $\ge 1-\delta$ over S, h_t:**

$$L_{\mathcal{D}}(h_t) - L_S(h_t) \le \sqrt{\frac{\beta + D_\infty(p_0\|\nu) + \ln {}^1/_\delta}{m}} \le \sqrt{\frac{\beta + \ln {}^1/_\delta}{m}}$$

for $0 \le loss \le 1$

p_0 not data dependent,
$\mathrm{d}p_\infty \propto e^{-\beta\mathcal{L}_S}\mathrm{d}p_0$, $\boldsymbol{\mathcal{L}_s \le 1}$

Application to Langevin Dynamics

$$\mathrm{d}h_t = -\nabla\mathcal{L}_S(h_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t$$

➔ $\mathrm{d}p_\infty \propto e^{-\beta\mathcal{L}_S}$

Application to Langevin Dynamics

- Reflective Langevin Dynamics on Bounded Domain:

$$\mathrm{d}h_t = -\nabla \mathcal{L}_S(h_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t + \mathrm{d}r_t$$

➔ $\mathrm{d}p_\infty \propto e^{-\beta \mathcal{L}_S}\mathrm{d}\nu$, $p_0 = \nu$ Uniform on box

- Regularized Langevin Dynamics:

$$\mathrm{d}h_t = -\nabla \mathcal{L}_S(h_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}W_t - \frac{\lambda}{\beta} h_t \mathrm{d}t$$

➔ $\mathrm{d}p_\infty \propto e^{-\beta \mathcal{L}_S}\mathrm{d}\nu$, $p_0 = \nu = \mathcal{N}(0, \lambda^{-1} I)$

"We are approaching AGI and it's not clear that knowing this tighter bound will get us closer to that" *–Reviewer 2 (of another paper)*

The journey is ~~half~~ the reward

In both cases: $\mathbb{E}_{h\sim p_t}[L_\mathcal{D}(h) - L_S(h)] \leq \sqrt{\frac{\beta \mathbb{E}_{p_0}\mathcal{L}_S + \ln{}^1/_\delta}{m}}$, with w.p.$\geq \delta$, $L_\mathcal{D}(h) \leq L_S(h) + \sqrt{\frac{\beta \sup \mathcal{L}_S + \ln{}^1/_\delta}{m}}$

Paper	Trajectory dependent	dimension dependence	Bound (big O)
Mou et al. [40]	✓	through gradients	$\sqrt{\frac{\beta}{N}} \cdot \sqrt{\frac{1}{\lambda} g_t^2}$
Li et al. [32]	✗	through K	$\frac{e^{4\beta C}\sqrt{\beta}}{N} \cdot \frac{2K}{\sqrt{\lambda}}$
Futami and Fujisawa [16]	✓	through gradients	$\sqrt{\frac{\beta}{N} e^{8\beta C}} \cdot \sqrt{\frac{1}{\lambda} g_t^2}$
Ours (11)	✗	✗	$\sqrt{\frac{\beta}{N}} \cdot \sqrt{C}$

K = Lip const

$\mathcal{L}_S \leq C$

Temperature is All You Need

for Generalization in Langevin Dynamics and other Markov Processes

Itamar Harel (Technion), Yonathan Walonowsky (Technion), Gal Vardi (Weizmann), Nati Srebro (TTIC), Daniel Soudry (Technion)

Second Law of Thermodynamics

$KL(p_t \| p_\infty) \le KL(p_0 \| p_\infty) \qquad D_\infty(p_t \| p_\infty) \le D_\infty(p_0 \| p_\infty)$

If p is Gibbs wrt q, i.e. $\mathrm{d}p \propto e^{-\Psi}\mathrm{d}q$ then

$$KL(p\|q) + KL(q\|p) = \mathbb{E}_q \Psi - \mathbb{E}_p \Psi$$

$$D_\infty(p\|q) + D_\infty(q\|p) = \sup_q \Psi - \inf_p \Psi$$

If exists stationary dist $\mathrm{d}p_\infty \propto e^{-\Psi}\mathrm{d}\nu,\ \Psi \ge 0$

$$KL(p_t\|\nu) \le KL(p_0\|\nu) + \mathbb{E}_{p_0}\Psi$$

$$D_\infty(p_t\|\nu) \le D_\infty(p_0\|\nu) + \sup \Psi$$

For any time-inv data-dependent Markov Process with some stationary distribution $p_\infty(\cdot; S)$ that is Gibbs w.r.t. a fixed (non data dependent) ν with potential $\Psi(h; S) \ge 0$, with prob $\ge 1-\delta$ over $S \sim \mathcal{D}^m$:

$$\mathbb{E}_{h_t}[L_\mathcal{D}(h_t) - L_S(h_t)] \le \sqrt{\frac{\mathbb{E}_{h\sim p_0}\Psi(h) + KL(p_0\|\nu) + \ln {}^1/_\delta}{m}} \le \sqrt{\frac{\beta + \ln {}^1/_\delta}{m}}$$

p_0 not data dependent, $\mathrm{d}p_\infty \propto e^{-\beta \mathcal{L}_S}\mathrm{d}p_0$, $\mathbb{E}_{p_0}\mathcal{L}_S \le 1$

And if $\Psi \le \beta$ then with prob $\ge 1-\delta$ over $S \sim \mathcal{D}^m$ AND h_t:

$$L_\mathcal{D}(h_t) - L_S(h_t) \le \sqrt{\frac{\beta + D_\infty(p_0\|\nu) + \ln {}^1/_\delta}{m}}$$

for $0 \le loss \le 1$