European Research Council

Established by the European Commission

Convergence of Adaptive Langevin dynamics

Gabriel STOLTZ

(CERMICS, Ecole des Ponts & MATHERIALS team, INRIA Paris)

In collaboration with B. Leimkuhler and M. Sachs

Work also supported by ANR Funding ANR-14-CE23-0012 ("COSMOS")

Outline

- **Adaptive Langevin dynamics**
  - Motivation: Bayesian inference for large data sets
- **Convergence of Langevin type dynamics**
  - A review of possible approaches
  - A focus on a "direct" hypocoercive technique
- **Convergence of Adaptive Langevin dynamics**
  - Heuristic demonstration of convergence rates
  - Sharp bounds through hypocoercive techniques
  - Central Limit Theorem and "Langevin" limit

Adaptive Langevin dynamics

Bayesian inference in the large data context

- **Data $\{x_i\}_{i=1,\dots,N}$ to be explained by a statistical model**
  - Parametrization by $q \in \mathbb{R}^n$: individual likelihoods $P(x_i|q)$
  - Prior $\rho(q)$ on the parameters
  - Sample q from $\nu(dq) = \mathrm{e}^{-V(q)}\, dq = Z_\nu^{-1} \rho(q) \prod_{i=1}^N P(x_i|q)\, dq$
  - For usual MCMC methods, each step costs $\mathrm{O}(N)$

- **Mini-batching**: Stochastic gradient Langevin dynamics[1]
  - Fundamental assumption: for $1 \ll \mathcal{N} \ll N$ and $J_{\mathcal{N}} \in \{1, \dots, N\}^{\mathcal{N}}$,

$$\nabla(\ln \rho)(q) + \frac{N}{\mathcal{N}} \sum_{j \in J_{\mathcal{N}}} \nabla(\ln P(x_j|q)) = -\nabla V(q) + \mathcal{G}, \quad \mathcal{G} \sim \mathcal{N}(0, \Sigma(q))$$

  - Amounts to introducing an additional Brownian motion of unknown magnitude $\rightarrow$ bias
  - Assume that $\Sigma(q)$ is constant [Work of Inass Sekkat...]

[1] Welling/Teh, *ICML* (2011)

Removing the mini-batching bias

- Phase-space extension: momenta p and variable friction ζ

Adaptive Langevin dynamics[1]: unknown σ (scalar, for simplicity)

$$
\begin{aligned}
dq &= M^{-1} p \, dt, \\
dp &= \left(-\nabla V(q) - \zeta M^{-1} p \right) dt + \sigma \, dW_t, \\
d\zeta &= \frac{1}{m} \left(p^T M^{-2} p - \beta^{-1} \mathrm{Tr}\left(M^{-1} \right) \right) dt
\end{aligned}
$$

- Invariant measure π with density proportional to

$$
\exp\left(-\beta \left[\frac{p^T M^{-1} p}{2} + V(q) + \frac{m}{2} \left(\zeta - \frac{\beta \sigma^2}{2} \right)^2 \right] \right) dq \, dp \, d\zeta
$$

- The marginal of π in q is indeed ν whatever σ... Prove convergence, in particular Central Limit Theorem?

[1] A. Jones and B. Leimkuhler, *J. Chem. Phys.* (2011); Ding et al., *NIPS* (2014); B. Leimkuhler and X. Shang, *SIAM J. Sci. Comput.* (2015)

Standard Langevin dynamics

Langevin dynamics

- hase-space $\mathcal{E} = \mathcal{D} \times \mathbb{R}^d$, Hamiltonian $H(q,p) = V(q) + \frac{1}{2} p^T M^{-1} p$

Stochastic perturbation of the Hamiltonian dynamics

$$\begin{cases} dq_t = M^{-1} p_t \, dt \\ \\ dp_t = -\nabla V(q_t) \, dt - \gamma M^{-1} p_t \, dt + \sqrt{\dfrac{2\gamma}{\beta}} \, dW_t \end{cases}$$

- Given (known) friction $\gamma > 0$ (could be a position-dependent matrix)

Generator of the Langevin dynamics $\mathcal{L} = \mathcal{L}_{\rm ham} + \gamma \mathcal{L}_{\rm FD}$

$$\mathcal{L}_{\rm ham} = p^T M^{-1} \nabla_q - \nabla V^T \nabla_p, \qquad \mathcal{L}_{\rm FD} = -p^T M^{-1} \nabla_p + \frac{1}{\beta} \Delta_p$$

- Invariant proba. measure $\mu(dq\,dp) = Z^{-1} \mathrm{e}^{-\beta H(q,p)} \, dq \, dp = \nu(dq) \, \kappa(dp)$

Ergodicity results (1)

- Almost-sure convergence[2] of ergodic averages $\widehat{\varphi}_t = \dfrac{1}{t}\displaystyle\int_0^t \varphi(q_s, p_s)\, ds$
- Asymptotic variance of ergodic averages

$$\sigma_\varphi^2 = \lim_{t\to+\infty} t\mathbb{E}\left[\widehat{\varphi}_t^2\right] = 2\int_{\mathcal{E}} \left(-\mathcal{L}^{-1}\Pi_0\varphi\right)\Pi_0\varphi\, d\mu$$

where $\Pi_0\varphi = \varphi - \mathbb{E}_\mu(\varphi)$

- A central limit theorem holds[3] when the equation has a solution in $L^2(\mu)$

Poisson equation in $L^2(\mu)$

$$-\mathcal{L}\Phi = \Pi_0\varphi$$

- Well-posedness of such equations?

---

[2] Kliemann, *Ann. Probab.* **15**(2), 690-707 (1987)

[3] Bhattacharya, *Z. Wahrsch. Verw. Gebiete* **60**, 185–201 (1982)

Ergodicity results (2)

- Invertibility of $\mathcal{L}$ on subsets of $L^2_0(\mu) = \left\{ \varphi \in L^2(\mu) \,\middle|\, \int_{\mathcal{E}} \varphi \, d\mu = 0 \right\}$?

$$-\mathcal{L}^{-1} = \int_0^{+\infty} \mathrm{e}^{t\mathcal{L}} \, dt$$

- Prove exponential convergence of the semigroup $\mathrm{e}^{t\mathcal{L}}$
  - various Banach spaces $E \cap L^2_0(\mu)$
  - Lyapunov techniques[4] $L^\infty_W(\mathcal{E}) = \left\{ \varphi \, \text{measurable}, \, \left\| \frac{\varphi}{W} \right\|_{L^\infty} < +\infty \right\}$
  - standard hypocoercive[5] setup $H^1(\mu)$
  - $E = L^2(\mu)$ after hypoelliptic regularization[6] from $H^1(\mu)$
  - coupling arguments[7]

---

[4] L. Wu, *Stoch. Proc. Appl.* (2001); Mattingly, Stuart and Higham, *Stoch. Proc. Appl.* (2002); L. Rey-Bellet, *Lect. Notes Math.* (2006); Hairer and Mattingly, *Progr. Probab.* (2011)

[5] Villani (2009) and before Talay (2002), Eckmann/Hairer (2003), Hérau/Nier (2004)

[6] F. Hérau, *J. Funct. Anal.* **244**(1), 95-118 (2007)

[7] A. Eberle, A. Guillin and R. Zimmer, *arXiv preprint* **1703.01617** (2017)

Direct $L^2(\mu)$ approach: lack of coercivity

- The generator, considered on $L^2(\mu)$, is the sum of...
  - a degenerate symmetric part $\mathcal{L}_{\mathrm{FD}} = -p^T M^{-1} \nabla_p + \frac{1}{\beta} \Delta_p$
  - an antisymmetric part $\mathcal{L}_{\mathrm{ham}} = p^T M^{-1} \nabla_q - \nabla V^T \nabla_p$

- Standard strategy for coercive generators: consider φ with average 0 with respect to μ and compute

$$\begin{aligned} \frac{d}{dt} \left(\left\| \mathrm{e}^{t\mathcal{L}} \varphi \right\|_{L^2(\mu)}^2 \right) &= \left\langle \mathrm{e}^{t\mathcal{L}} \varphi, \mathcal{L} \mathrm{e}^{t\mathcal{L}} \right\rangle_{L^2(\mu)} = \left\langle \mathrm{e}^{t\mathcal{L}} \varphi, \mathcal{L}_{\mathrm{FD}} \mathrm{e}^{t\mathcal{L}} \right\rangle_{L^2(\mu)} \\ &= -\frac{1}{\beta} \left\| \nabla_p \mathrm{e}^{t\mathcal{L}} \varphi \right\|_{L^2(\mu)}^2 \leqslant 0, \end{aligned}$$

but no control of $\|\phi\|_{L^2(\mu)}$ by $\|\nabla_p \phi\|_{L^2(\mu)}$ for a Gronwall estimate...

- Change of scalar product in order to use the antisymmetric part

Almost direct $L^2(\mu)$ approach: convergence result

- Assume that the potential V is smooth and[8,9]
  - the marginal measure ν satisfies a Poincaré inequality
$$\|\Pi_0 \varphi\|^2_{L^2(\nu)} \leqslant \frac{1}{C_\nu} \|\nabla_q \varphi\|^2_{L^2(\nu)}$$
  - there exist $c_1 > 0$, $c_2 \in [0,1)$ and $c_3 > 0$ such that V satisfies
$$\Delta V \leqslant c_1 + \frac{c_2}{2} |\nabla V|^2, \quad |\nabla^2 V| \leqslant c_3 \left(1 + |\nabla V|\right)$$

There exist $C > 0$ and $\lambda_\gamma > 0$ such that, for any $\varphi \in L^2_0(\mu)$,

$$\forall t \geqslant 0, \qquad \left\| \mathrm{e}^{t\mathcal{L}} \varphi \right\|_{L^2(\mu)} \leqslant C \mathrm{e}^{-\lambda_\gamma t} \|\varphi\|_{L^2(\mu)}.$$

with convergence rate of order $\min(\gamma, \gamma^{-1})$: there exists $\overline{\lambda} > 0$ such that

$$\lambda_\gamma \geqslant \overline{\lambda} \min(\gamma, \gamma^{-1}).$$

[8] Dolbeault, Mouhot and Schmeiser, *C. R. Math. Acad. Sci. Paris* (2009)
[9] Dolbeault, Mouhot and Schmeiser, *Trans. AMS*, **367**, 3807–3828 (2015)

Sketch of proof

- Modified square norm $\mathcal{H}[\varphi] = \frac{1}{2}\|\varphi\|^2 - \varepsilon \langle A\varphi, \varphi \rangle$ for $\varepsilon \in (-1, 1)$ and

$$A = \left(1 + (\mathcal{L}_{\mathrm{ham}}\Pi_p)^*(\mathcal{L}_{\mathrm{ham}}\Pi_p)\right)^{-1}(\mathcal{L}_{\mathrm{ham}}\Pi_p)^*, \qquad \Pi_p\varphi = \int_{\mathbb{R}^D} \varphi \, d\kappa$$

- $A = \Pi_p A(1 - \Pi_p)$ and $\mathcal{L}_{\mathrm{ham}}A$ are bounded so that $\mathcal{H} \sim \|\cdot\|^2_{L^2(\mu)}$

Coercivity in the scalar product $\langle\langle \cdot, \cdot \rangle\rangle$ induced by $\mathcal{H}$

$$\mathscr{D}[\varphi] := \langle\langle -\mathcal{L}\varphi, \varphi \rangle\rangle \geqslant \widetilde{\lambda}_\gamma \|\varphi\|^2$$

- Idea: control of $\|(1 - \Pi_p)\varphi\|^2$ by $\langle -\mathcal{L}_{\mathrm{FD}}\varphi, \varphi \rangle$ (Poincaré); for $\|\Pi_p\varphi\|^2$,

$$\|\mathcal{L}_{\mathrm{ham}}\Pi_p\varphi\|^2 \geqslant \frac{DC_\nu}{\beta m}\|\Pi_p\varphi\|^2, \qquad \text{hence } A\mathcal{L}_{\mathrm{ham}}\Pi_p \geqslant \lambda_{\mathrm{ham}}\Pi_p$$

- Gronwall inequality $\frac{d}{dt}\left(\mathcal{H}\left[\mathrm{e}^{t\mathcal{L}}\varphi\right]\right) = -\mathscr{D}\left[\mathrm{e}^{t\mathcal{L}}\varphi\right] \leqslant -\frac{2\widetilde{\lambda}_\gamma}{1+\varepsilon}\mathcal{H}\left[\mathrm{e}^{t\mathcal{L}}\varphi\right]$

Convergence of Adaptive Langevin dynamics

Structure of Adaptive Langevin dynamics (1)

- Normalization of the dynamics, for the invariant measure to be independent of m (take $M = \mathrm{Id}$ to simplify)

$$\begin{cases} dq_t = p_t\, dt, \\ dp_t = (-\nabla V(q_t) - \zeta_t p_t)\, dt + \sigma\, dW_t, \\ d\zeta_t = \dfrac{1}{m}\left(|p_t|^2 - \dfrac{n}{\beta}\right) dt \end{cases}$$

- Set $\varepsilon = \sqrt{m}$ and $\zeta = \gamma + \dfrac{\xi}{\varepsilon}$ with $\gamma = \beta\sigma^2/2$

Normalized Adaptive Langevin dynamics

$$\begin{cases} dq_t = p_t\, dt, \\ dp_t = \left(-\nabla V(q_t) - \dfrac{\xi_t}{\varepsilon} p_t - \gamma p_t\right) dt + \sqrt{\dfrac{2\gamma}{\beta}}\, dW_t, \\ d\xi_t = \dfrac{1}{\varepsilon}\left(|p_t|^2 - \dfrac{n}{\beta}\right) dt \end{cases}$$

Structure of Adaptive Langevin dynamics (2)

- Invariant measure π with density $Z^{-1} \exp\left(-\beta\left[\frac{|p|^2}{2} + V(q) + \frac{\xi^2}{2}\right]\right)$
- Generator $\mathcal{L}_{\mathrm{AdL}} = \mathcal{L}_{\mathrm{ham}} + \gamma\mathcal{L}_{\mathrm{FD}} + \varepsilon^{-1}\mathcal{L}_{\mathrm{NH}}$ with (adjoints on $L^2(\pi)$)

$$\mathcal{L}_{\mathrm{ham}} = \frac{1}{\beta}\left(\nabla_p^*\nabla_q - \nabla_q^*\nabla_p\right) = \frac{1}{\beta}\sum_{i=1}^{n} \partial_{p_i}^*\partial_{q_i} - \partial_{q_i}^*\partial_{p_i},$$

$$\mathcal{L}_{\mathrm{FD}} = -\frac{1}{\beta}\nabla_p^*\nabla_p = -\frac{1}{\beta}\sum_{i=1}^{n} \partial_{p_i}^*\partial_{p_i},$$

$$\mathcal{L}_{\mathrm{NH}} = \left(|p|^2 - \frac{n}{\beta}\right)\partial_\xi - \xi\, p^T\nabla_p = \frac{1}{\beta^2}\left((\partial_\xi - \partial_\xi^*)\nabla_p^*\nabla_p + \Delta_p^*\partial_\xi - \Delta_p\partial_\xi^*\right)$$

- Antisymmetric parts $\mathcal{L}_{\mathrm{ham}}$, $\mathcal{L}_{\mathrm{NH}}$ and symmetric one $\mathcal{L}_{\mathrm{FD}}$
- Proof of exponential decay using Lyapunov techniques[10]

[10] D. Herzog, *Commun. Math. Sci.* (2018)

Expected scalings

- Generator $\simeq$ superposition of $\mathcal{L}_{\mathrm{ham}} + \gamma\mathcal{L}_{\mathrm{FD}}$ and $\varepsilon^{-1}\mathcal{L}_{\mathrm{NH}} + \gamma\mathcal{L}_{\mathrm{FD}}$
  - Exponential rate of decay $\sim \min(\gamma, \gamma^{-1})$ for the Langevin part
  - Nosé–Hoover-like part rewritten as $\varepsilon^{-1}(\mathcal{L}_{\mathrm{NH}} + \gamma\varepsilon\mathcal{L}_{\mathrm{FD}})$
    $\rightarrow$ suggests rate of decay $\sim \varepsilon^{-1}\min(\gamma\varepsilon, (\gamma\varepsilon)^{-1})$

Exponential convergence of the semigroup

There exist $C, \overline{\lambda}$ such that, for any $\varepsilon, \gamma > 0$, there is $\lambda_{\varepsilon,\gamma} > 0$ for which

$$\forall t \geqslant 0,\ \forall \varphi \in L^2(\pi), \quad \left\| \mathrm{e}^{t\mathcal{L}_{\mathrm{AdL}}}\varphi - \int \varphi \, d\pi \right\|_{L^2(\pi)} \leqslant C\mathrm{e}^{-\lambda_{\varepsilon,\gamma}} \left\| \varphi - \int \varphi \, d\pi \right\|_{L^2(\pi)}$$

with the lower bound $\lambda_{\varepsilon,\gamma} \geqslant \overline{\lambda}\min\left(\gamma, \frac{1}{\gamma}, \frac{1}{\gamma\varepsilon^2}\right)$. As a consequence,

$$\mathcal{L}_{\mathrm{AdL}}^{-1} = -\int_0^\infty \mathrm{e}^{t\mathcal{L}_{\mathrm{AdL}}}\, dt, \qquad \left\|\mathcal{L}_{\mathrm{AdL}}^{-1}\right\|_{\mathcal{B}(L^2_0(\pi))} \leqslant \frac{C}{\overline{\lambda}} \max\left(\gamma, \gamma^{-1}, \gamma\varepsilon^2\right).$$

Sharpness of the scaling and elements of proof

- Scaling of resolvent norm sharp in view of specific solutions, *e.g.*

$$\mathcal{L}_{\mathrm{AdL}}\left(\gamma\varepsilon\xi+\frac{|p|^2}{2}-\frac{p^T\nabla V}{\gamma}\right)=-\frac{\xi|p|^2}{\varepsilon}+\frac{p^T\nabla V}{\gamma\varepsilon}-\frac{1}{\gamma}\left(p^T\nabla^2 V p-|\nabla V|^2\right),$$

which shows that $\left\|\mathcal{L}_{\mathrm{AdL}}^{-1}\right\|_{\mathcal{B}(L^2_0(\pi))}\geqslant c\gamma\varepsilon^2$ by choosing $\gamma\gg\varepsilon\gg 1$

- **Proof:** construction of regularization operator with correct scaling
  - total antisymmetric part $\mathscr{A}_\varepsilon=\mathcal{L}_{\mathrm{ham}}+\varepsilon^{-1}\mathcal{L}_{\mathrm{NH}}$
  - distinguish $\varepsilon\leqslant 1$ (Langevin limits convergence) or $\varepsilon\geqslant 1$

$$A_\varepsilon:=-\min\left(1,\frac{1}{\varepsilon}\right)\left[\min\left(1,\frac{1}{\varepsilon^2}\right)-\Pi\mathscr{A}_\varepsilon^2\Pi\right]^{-1}\Pi\mathscr{A}_\varepsilon$$

$$=-\min\left(1,\frac{1}{\varepsilon}\right)\left[\min\left(1,\frac{1}{\varepsilon^2}\right)+\Pi\left(\frac{2n}{(\beta\varepsilon)^2}\partial_\xi^*\partial_\xi+\frac{1}{\beta}\nabla_q^*\nabla_q\right)\Pi\right]^{-1}\Pi\mathscr{A}_\varepsilon$$

  - technical estimates to control all terms uniformly in ε

Central Limit Theorem

- Consider $\varphi \in L^2(\pi)$ and $\overline{\varphi}_t := \dfrac{1}{t}\displaystyle\int_0^t \varphi(q_s, p_s, \xi_s)\, ds$

Central Limit Theorem

$$\sqrt{t}\,(\widehat{\varphi}_t - \mathbb{E}_\pi \varphi) \xrightarrow[t\to+\infty]{\text{law}} \mathcal{N}(0, \sigma^2_{\varepsilon,\gamma}(\varphi)),$$

with the asymptotic variance (with $\Pi_0 \varphi = \varphi - \mathbb{E}_\pi(\varphi)$)

$$0 \leqslant \sigma^2_{\varepsilon,\gamma}(\varphi) = 2\int \left(-\mathcal{L}^{-1}_{\mathrm{AdL}} \Pi_0 \varphi\right) \Pi_0 \varphi \, d\pi \leqslant 2C\overline{\lambda}^{-1} \|\varphi\|^2_{L^2(\pi)} \max\left(\gamma, \gamma^{-1}, \gamma\varepsilon^2\right)$$

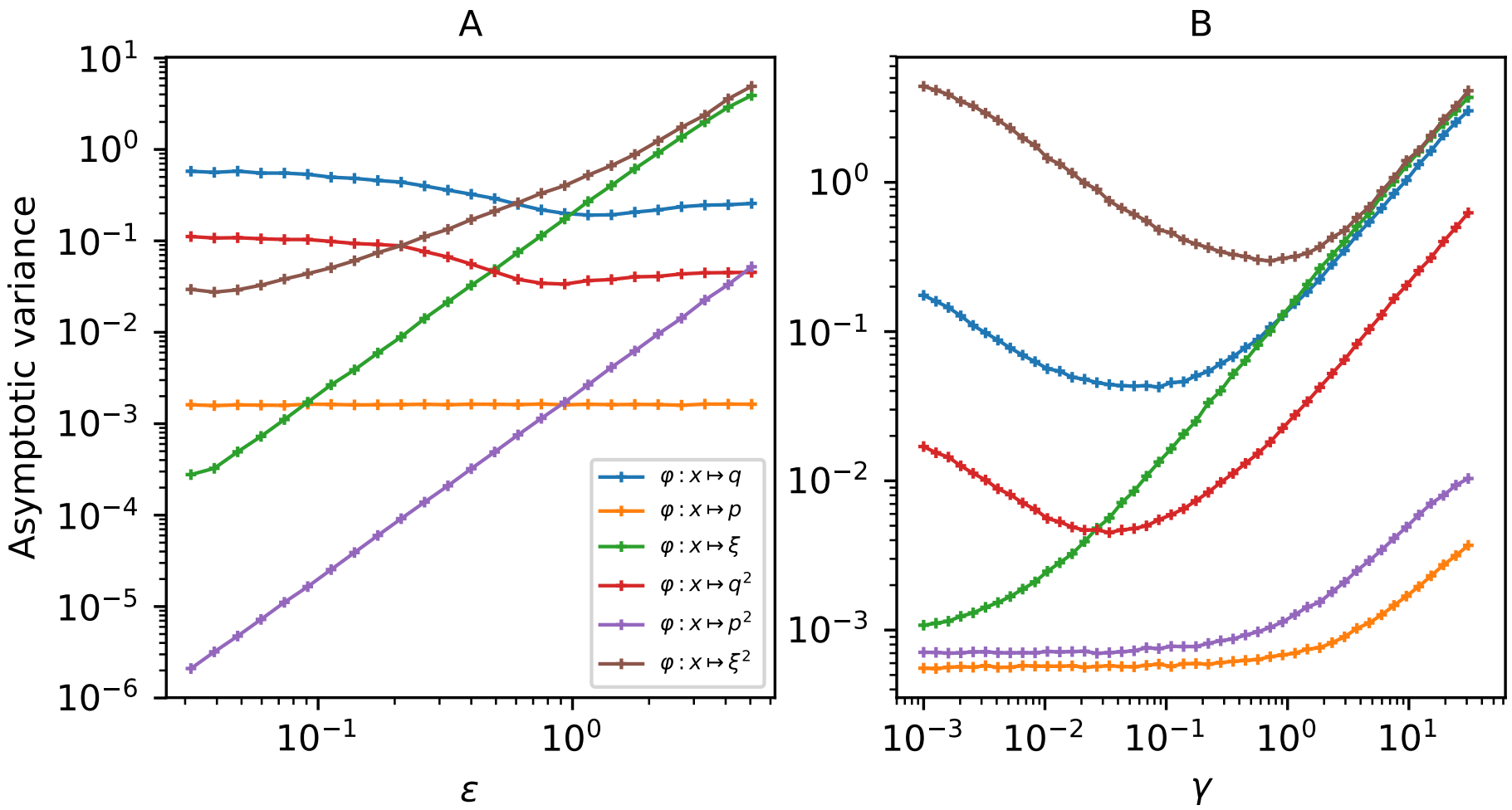
- **Suggests taking** $\gamma = 1$ **and** $\varepsilon \to 0$ (beware numerical stability)
- Langevin type limit $\varepsilon \to +\infty$ for a function $\varphi(q, p)$ (independent of ξ)

$$\left|\sigma^2_{\varepsilon,\gamma}(\varphi) - \sigma^2_{\infty,\gamma}(\varphi)\right| \leqslant \frac{K}{\varepsilon}$$

Proof: asymptotic analysis and fine estimates[11] of $\mathcal{L}_{\mathrm{Lang}} = \mathcal{L}_{\mathrm{ham}} + \gamma \mathcal{L}_{\mathrm{FD}}$

[11] Talay, *Markov Proc. Rel. Fields* (2002); Kopec, *BIT* (2015)

Some numerical results



Letft: scaling $\max(1, \varepsilon^2)$ of the variance (γ fixed).
Right: scaling $\max(\gamma, \gamma^{-1})$ of the variance (ε fixed).