A Mean-Field Theory of Lazy Training in Two-Layer Neural Nets: Entropic Regularization and Controlled McKean–Vlasov Dynamics

Maxim Raginsky

(joint work with **Belinda Tzen**)

University of Illinois at Urbana-Champaign

April 2020

Motivation

Infinitely wide neural nets

- mean-field view: abstract away finite-size effects, focus on universal phenomena
- training dynamics: PDE on the space of probability measures (distributional or nonlinear dynamics)
- finite-neuron setting can be obtained by sampling/discretization

Effects of massive overparametrization and hyperparameter choices

- *lazy training* (Chizat–Oyallon–Bach, 2019): weights barely move from random Gaussian initialization, while risk decreases
- *Neural Tangent Kernel* (Jacot–Gabriel–Hongler, 2018): during training, neural net stays close to its lineraization around initialization

In this talk: an alternative mean-field theory of lazy training via *entropic regularization*.

Related Work

Related Work

- Mean-field limit:

  Nitanda–Suzuki, 2017; Chizat and Bach, 2018;
  Mei–Montanari–Nguyen, 2018; Rotskoff and Vanden–Eijnden, 2018;
  Mei–Misiakiewicz–Montanari, 2019; Sirignano and Spiliopoulos, 2020

Related Work

- Mean-field limit:

  Nitanda–Suzuki, 2017; Chizat and Bach, 2018;
  Mei–Montanari–Nguyen, 2018; Rotskoff and Vanden–Eijnden, 2018;
  Mei–Misiakiewicz–Montanari, 2019; Sirignano and Spiliopoulos, 2020

- Lazy training and the Neural Tangent Kernel:

  Jacot–Gabriel–Hongler, 2018; Allen-Zhu–Li–Liang, 2019;
  Chizat–Oyallon–Bach, 2019

Related Work

- Mean-field limit:

  Nitanda–Suzuki, 2017; Chizat and Bach, 2018;
  Mei–Montanari–Nguyen, 2018; Rotskoff and Vanden–Eijnden, 2018;
  Mei–Misiakiewicz–Montanari, 2019; Sirignano and Spiliopoulos, 2020

- Lazy training and the Neural Tangent Kernel:

  Jacot–Gabriel–Hongler, 2018; Allen-Zhu–Li–Liang, 2019;
  Chizat–Oyallon–Bach, 2019

- Universal approximation by transporting weights:

  Ji–Telgarsky–Xian, 2020

Function Approximation by Two-Layer Neural Nets

- Two-layer net with N hidden units:

$$\hat{f}_N(x; \boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; w^i), \qquad w^i \in \mathbb{R}^d$$

Function Approximation by Two-Layer Neural Nets

- Two-layer net with N hidden units:

$$\hat{f}_N(x; \boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; w^i), \qquad w^i \in \mathbb{R}^d$$

- Approximation risk for a target function $f : \mathcal{X} \to \mathbb{R}$:

$$R_N(\boldsymbol{w}) := \|f - \hat{f}_N(\cdot; \boldsymbol{w})\|^2_{L^2(\pi)} = \int_{\mathcal{X}} \pi(\mathrm{d}x)(f(x) - \hat{f}_N(x; \boldsymbol{w}))^2$$

Function Approximation by Two-Layer Neural Nets

- Two-layer net with N hidden units:

$$\hat{f}_N(x; \boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; w^i), \qquad w^i \in \mathbb{R}^d$$

- Approximation risk for a target function $f : \mathcal{X} \to \mathbb{R}$:

$$R_N(\boldsymbol{w}) := \|f - \hat{f}_N(\cdot; \boldsymbol{w})\|^2_{L^2(\pi)} = \int_{\mathcal{X}} \pi(\mathrm{d}x)(f(x) - \hat{f}_N(x; \boldsymbol{w}))^2$$

- The risk depends only on the empirical distribution of $\boldsymbol{w}$:

$$R_N(\boldsymbol{w}) = R_0 + \frac{2}{N} \sum_{i=1}^{N} \tilde{f}(w^i) + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K(w^i, w^j)$$

$$\begin{aligned} \text{where } R_0 &:= \mathbf{E}_\pi[f^2(X)], \\ \tilde{f}(w) &:= -\mathbf{E}_\pi[f(X)\sigma(X; w)], \\ K(w, \tilde{w}) &:= \mathbf{E}_\pi[\sigma(X; w)\sigma(X; \tilde{w})] \end{aligned}$$

Mean-Field Limit: A Continuum of Neurons

- From finite population of neurons to a continual ensemble:

$$\hat{\mu}_{\boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{w^i} \quad \longrightarrow \quad \mu \in \mathscr{P}(\mathbb{R}^d)$$

$$\hat{f}_N(x;\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\sigma(x;w^i) \quad \longrightarrow \quad \hat{f}(x;\mu) = \int_{\mathbb{R}^d}\mu(\mathrm{d}w)\sigma(x;w)$$

Mean-Field Limit: A Continuum of Neurons

- From finite population of neurons to a continual ensemble:

$$\hat{\mu}_{\boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{w^i} \quad \longrightarrow \quad \mu \in \mathscr{P}(\mathbb{R}^d)$$

$$\hat{f}_N(x;\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\sigma(x;w^i) \quad \longrightarrow \quad \hat{f}(x;\mu) = \int_{\mathbb{R}^d}\mu(\mathrm{d}w)\sigma(x;w)$$

- L^2 risk:

$$R(\mu) = \|f - \hat{f}(\cdot;\mu)\|^2_{L^2(\pi)} = R_0 + 2\int \tilde{f}\,\mathrm{d}\mu + \int K\,\mathrm{d}(\mu\otimes\mu)$$

— *convex quadratic* in the measure μ

Mean-Field Limit: A Continuum of Neurons

- From finite population of neurons to a continual ensemble:

$$\hat{\mu}_{\boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{w^i} \quad \longrightarrow \quad \mu \in \mathscr{P}(\mathbb{R}^d)$$

$$\hat{f}_N(x;\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\sigma(x;w^i) \quad \longrightarrow \quad \hat{f}(x;\mu) = \int_{\mathbb{R}^d}\mu(\mathrm{d}w)\sigma(x;w)$$

- L^2 risk:

$$R(\mu) = \|f - \hat{f}(\cdot;\mu)\|^2_{L^2(\pi)} = R_0 + 2\int \tilde{f}\,\mathrm{d}\mu + \int K\,\mathrm{d}(\mu\otimes\mu)$$

  — *convex quadratic* in the measure μ

- Back to finite N: $R_N(\boldsymbol{w}) = R(\hat{\mu}_{\boldsymbol{w}})$

Entropy-Regularized Risk in the Mean-Field Limit

Free energy:

$$F_{\beta,\tau}(\mu) := \frac{1}{2}R(\mu) + \frac{\tau}{\beta}D_{\mathrm{KL}}(\mu\|\gamma_\tau)$$

where $\gamma_\tau = \mathcal{N}(0, \tau I_d)$ is a Gaussian "prior" on the weights

Entropy-Regularized Risk in the Mean-Field Limit

Free energy:

$$F_{\beta,\tau}(\mu) := \frac{1}{2}R(\mu) + \frac{\tau}{\beta}D_{\mathrm{KL}}(\mu\|\gamma_\tau)$$

where $\gamma_\tau = \mathcal{N}(0, \tau I_d)$ is a Gaussian "prior" on the weights

Some motivation:

Entropy-Regularized Risk in the Mean-Field Limit

Free energy:

$$F_{\beta,\tau}(\mu) := \frac{1}{2}R(\mu) + \frac{\tau}{\beta}D_{\mathrm{KL}}(\mu\|\gamma_\tau)$$

where $\gamma_\tau = \mathcal{N}(0, \tau I_d)$ is a Gaussian "prior" on the weights

Some motivation:

- If $\mu = \mathcal{N}(\theta, \tau I_d)$, then

$$F_{\beta,\tau}(\mu) = \frac{1}{2}R(\mu) + \frac{1}{2\beta}\|\theta\|_2^2$$

  — a form of ridge regression

Entropy-Regularized Risk in the Mean-Field Limit

Free energy:

$$F_{\beta,\tau}(\mu) := \frac{1}{2}R(\mu) + \frac{\tau}{\beta}D_{\mathrm{KL}}(\mu\|\gamma_\tau)$$

where $\gamma_\tau = \mathcal{N}(0, \tau I_d)$ is a Gaussian "prior" on the weights

Some motivation:

- If $\mu = \mathcal{N}(\theta, \tau I_d)$, then

$$F_{\beta,\tau}(\mu) = \frac{1}{2}R(\mu) + \frac{1}{2\beta}\|\theta\|_2^2$$

  — a form of ridge regression

- Talagrand's entropy-transport inequality

$$D_{\mathrm{KL}}(\mu\|\gamma_\tau) \geq \frac{1}{2\tau}\mathcal{W}_2^2(\mu, \gamma_\tau),$$

  where $\mathcal{W}_2(\mu,\nu) := \min_{X\sim\mu, Y\sim\nu} \sqrt{\mathbf{E}[\|X-Y\|_2^2]}$, gives

$$F_{\beta,\tau}(\mu) \geq \frac{1}{2}R(\mu) + \frac{1}{2\beta}\mathcal{W}_2^2(\mu, \gamma_\tau)$$

Entropy-Regularized Risk: Static Formulation

Assumptions:

(i) f, σ bounded; (ii) $\nabla_w \sigma(X; w)$ subgaussian ($X \sim \pi$); (iii) $\tilde{f}, K$ Lipschitz with Lipschitz gradients

Theorem (Tzen–Raginsky, 2020) The free energy $F_{\beta,\tau}(\cdot)$ admits a *unique* minimizer $\mu^\star = \mu^\star_{\beta,\tau}$, such that:

1. The Boltzmann fixed-point condition holds:

$$\frac{\mathrm{d}\mu^\star}{\mathrm{d}\gamma_\tau}(w) \propto \exp\left(-\frac{\beta}{\tau}\Psi(w; \mu^\star)\right)$$

   where $\Psi(w; \mu) := \tilde{f}(w) + \int K(w, \tilde{w})\mu(\mathrm{d}\tilde{w})$

2. The optimal weight distribution is nearly Gaussian:

$$D_{\mathrm{KL}}(\mu^\star \| \gamma_\tau) \leq \frac{\kappa\beta^2}{\tau}, \qquad \mathcal{W}_2^2(\mu^\star, \gamma_\tau) \leq \kappa\beta^2$$

3. If $f = \hat{f}(\cdot; \mu^\circ)$, then $R(\mu^\star) \leq \frac{2\tau}{\beta} D_{\mathrm{KL}}(\mu^\circ \| \gamma_\tau)$.

Some Consequences

- Weak regularization: $\beta \to \infty$ while $\tau = o(\beta)$ gives

$$\lim_{\beta \uparrow \infty} \inf_{\mu} F_{\beta,\tau}(\mu) = \frac{1}{2} \inf_{\mu} R(\mu)$$

Some Consequences

- Weak regularization: $\beta \to \infty$ while $\tau = o(\beta)$ gives

$$\lim_{\beta \uparrow \infty} \inf_{\mu} F_{\beta,\tau}(\mu) = \frac{1}{2} \inf_{\mu} R(\mu)$$

- Strong regularization: both β and τ are small but $\tau \ll \beta$:

$$\tau = \varepsilon^2/d, \;\; \beta = \sqrt{\tau d} = \varepsilon$$
$$D_{\mathrm{KL}}(\mu^\star \| \gamma_\tau) \le \kappa d, \quad \mathcal{W}_2^2(\mu^\star, \gamma_\tau) \le \kappa \varepsilon$$

Some Consequences

- Weak regularization: $\beta \to \infty$ while $\tau = o(\beta)$ gives

$$\lim_{\beta \uparrow \infty} \inf_{\mu} F_{\beta,\tau}(\mu) = \frac{1}{2} \inf_{\mu} R(\mu)$$

- Strong regularization: both β and τ are small but $\tau \ll \beta$:

$$\tau = \varepsilon^2/d, \;\; \beta = \sqrt{\tau d} = \varepsilon$$
$$D_{\mathrm{KL}}(\mu^\star \| \gamma_\tau) \le \kappa d, \quad \mathcal{W}_2^2(\mu^\star, \gamma_\tau) \le \kappa \varepsilon$$

- Realizable case: if $f = \hat{f}(\cdot; \mu^\circ)$ with $D_{\mathrm{KL}}(\mu^\circ \| \gamma_\tau) = O(d)$, then

$$R(\mu^\star) \le \frac{2\tau}{\beta} D_{\mathrm{KL}}(\mu^\circ \| \gamma_\tau) = \frac{2\kappa\varepsilon^2/d}{\varepsilon} \cdot O(d) = O(\varepsilon)$$

Lazy Training?

- Consider random Gaussian initialization:

$$\boldsymbol{W} = (W^1, \dots, W^N) \sim (\gamma_\tau)^{\otimes N} \quad \longrightarrow \quad \hat{f}_N(x; \boldsymbol{W}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; W^i)$$

Lazy Training?

- Consider random Gaussian initialization:

$$\boldsymbol{W} = (W^1, \ldots, W^N) \sim (\gamma_\tau)^{\otimes N} \quad \longrightarrow \quad \hat{f}_N(x; \boldsymbol{W}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; W^i)$$

- With high probability, there exist good approximations to "low-complexity" f near $\hat{f}_N(\cdot; \boldsymbol{W})$ (Allen-Zhu–Li–Liang, 2019)

Lazy Training?

- Consider random Gaussian initialization:

$$\boldsymbol{W} = (W^1, \ldots, W^N) \sim (\gamma_\tau)^{\otimes N} \quad \longrightarrow \quad \hat{f}_N(x; \boldsymbol{W}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; W^i)$$

- With high probability, there exist good approximations to "low-complexity" f near $\hat{f}_N(\cdot; \boldsymbol{W})$ (Allen-Zhu–Li–Liang, 2019)
- These good approximations can be found by gradient descent, and the weight updates do not drift too far from $\boldsymbol{W}$

Lazy Training?

- Consider random Gaussian initialization:

$$\boldsymbol{W} = (W^1, \ldots, W^N) \sim (\gamma_\tau)^{\otimes N} \quad \longrightarrow \quad \hat{f}_N(x; \boldsymbol{W}) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x; W^i)$$

- With high probability, there exist good approximations to "low-complexity" f near $\hat{f}_N(\cdot; \boldsymbol{W})$ (Allen-Zhu–Li–Liang, 2019)
- These good approximations can be found by gradient descent, and the weight updates do not drift too far from $\boldsymbol{W}$
- Transport map interpretation: $W^i \mapsto T(W^i)$ for some $T : \mathbb{R}^d \to \mathbb{R}^d$ (Ji–Telgarsky–Xian, 2020):

$$\hat{f}_N(\cdot; \boldsymbol{W}) \longmapsto \hat{f}_N(\cdot; T(\boldsymbol{W}))$$

  such that $\|T(W^i) - W^i\|_2$ is small

Corollary (Tzen–Raginsky, 2020) For β sufficiently small, there exists a Lipschitz transport map $T : \mathbb{R}^d \to \mathbb{R}^d$, such that all of the following holds with probability at least $1 - \delta$ for $\boldsymbol{W} = (W^1, \ldots, W^N) \sim (\gamma_\tau)^{\otimes N}$:

1. The transported weights $T(W^i)$ are uniformly close to the i.i.d. Gaussian weights W^i:

$$\max_{i \in [N]} \|T(W^i) - W^i\|_2 \leq \kappa\beta + \kappa\sqrt{\tau(\log N + \log(1/\delta))};$$

2. The transported neural net $\hat{f}_N(\cdot; T(\boldsymbol{W}))$ satisfies

$$\|f - \hat{f}_N(\cdot; T(\boldsymbol{W}))\|_{L^2(\pi)} \leq \|f - \hat{f}(\cdot; \mu^\star_{\beta,\tau})\|_{L^2(\pi)} + \kappa\sqrt{\frac{\log(1/\delta)}{N}};$$

3. The transported neural net $\hat{f}_N(\cdot; T(\boldsymbol{W}))$ and the random Gaussian neural net $\hat{f}_N(\cdot; \boldsymbol{W})$ are close in $L^2(\pi)$ norm:

$$\left\|\hat{f}_N(\cdot; T(\boldsymbol{W})) - \hat{f}_N(\cdot; \boldsymbol{W})\right\|^2_{L^2(\pi)} \leq \kappa\beta + \kappa\sqrt{\frac{\tau \log(1/\delta)}{N}}.$$

The Proof: Some Intuition

- Optimal $\mu^\star$ (Boltzmann fixed-point condition):

$$\mu^\star(\mathrm{d}w) \propto \exp\left(-\frac{1}{2\tau}\|w\|_2^2 - \frac{\beta}{\tau}\Psi(w;\mu^\star)\right)\mathrm{d}w$$

The Proof: Some Intuition

- Optimal $\mu^\star$ (Boltzmann fixed-point condition):

$$\mu^\star(\mathrm{d}w) \propto \exp\left(-\frac{1}{2\tau}\|w\|_2^2 - \frac{\beta}{\tau}\Psi(w;\mu^\star)\right)\mathrm{d}w$$

- Since $\nabla\Psi$ is Lipschitz, $w \mapsto \frac{1}{2}\|w\|_2^2 + \beta\Psi(w;\mu^\star)$ is strongly convex for all $\beta < \beta_0$; thus, the optimal (Brenier–McCann) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that achieves $\mathcal{W}_2(\mu^\star, \gamma_\tau)$ is Lipschitz by Caffarelli's regularity theorem.

The Proof: Some Intuition

- Optimal $\mu^\star$ (Boltzmann fixed-point condition):

$$\mu^\star(\mathrm{d}w) \propto \exp\left(-\frac{1}{2\tau}\|w\|_2^2 - \frac{\beta}{\tau}\Psi(w;\mu^\star)\right)\mathrm{d}w$$

- Since $\nabla\Psi$ is Lipschitz, $w \mapsto \frac{1}{2}\|w\|_2^2 + \beta\Psi(w;\mu^\star)$ is strongly convex for all $\beta < \beta_0$; thus, the optimal (Brenier–McCann) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that achieves $\mathcal{W}_2(\mu^\star, \gamma_\tau)$ is Lipschitz by Caffarelli's regularity theorem.
- Key consequences:

The Proof: Some Intuition

- Optimal $\mu^\star$ (Boltzmann fixed-point condition):

$$\mu^\star(\mathrm{d}w) \propto \exp\left(-\frac{1}{2\tau}\|w\|_2^2 - \frac{\beta}{\tau}\Psi(w;\mu^\star)\right)\mathrm{d}w$$

- Since $\nabla\Psi$ is Lipschitz, $w \mapsto \frac{1}{2}\|w\|_2^2 + \beta\Psi(w;\mu^\star)$ is strongly convex for all $\beta < \beta_0$; thus, the optimal (Brenier–McCann) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that achieves $\mathcal{W}_2(\mu^\star, \gamma_\tau)$ is Lipschitz by Caffarelli's regularity theorem.
- Key consequences:
  - can sample $W^1, \ldots, W^N \overset{\text{i.i.d.}}{\sim} \gamma_\tau$, then transport to $\mu^\star$:

$$\boldsymbol{W} \mapsto T(\boldsymbol{W}) = (T(W^1), \ldots, T(W^N)) \sim (\mu^\star)^{\otimes N};$$

    high-prob. guarantee for L^2 risk via refined Maurey's empirical method (Ji–Telgarsky–Xian, 2020)

The Proof: Some Intuition

- Optimal $\mu^\star$ (Boltzmann fixed-point condition):

$$\mu^\star(\mathrm{d}w) \propto \exp\left(-\frac{1}{2\tau}\|w\|_2^2 - \frac{\beta}{\tau}\Psi(w;\mu^\star)\right)\mathrm{d}w$$

- Since $\nabla\Psi$ is Lipschitz, $w \mapsto \frac{1}{2}\|w\|_2^2 + \beta\Psi(w;\mu^\star)$ is strongly convex for all $\beta < \beta_0$; thus, the optimal (Brenier–McCann) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that achieves $\mathcal{W}_2(\mu^\star, \gamma_\tau)$ is Lipschitz by Caffarelli's regularity theorem.
- Key consequences:
  - can sample $W^1, \ldots, W^N \overset{\text{i.i.d.}}{\sim} \gamma_\tau$, then transport to $\mu^\star$:

$$\boldsymbol{W} \mapsto T(\boldsymbol{W}) = (T(W^1), \ldots, T(W^N)) \sim (\mu^\star)^{\otimes N};$$

    high-prob. guarantee for L^2 risk via refined Maurey's empirical method (Ji–Telgarsky–Xian, 2020)
  - the functions

$$w \mapsto \|T(w) - w\|_2 \quad \text{and} \quad \boldsymbol{w} \mapsto \|\hat{f}_N(\cdot;\boldsymbol{w}) - \hat{f}_N(\cdot;T(\boldsymbol{w}))\|_2^2$$

    are Lipschitz, so we can use Gaussian concentration inequalities

Entropy-Regularized Risk: Dynamic Formulation

In a nutshell: we will construct a flow of measures $\boldsymbol{\mu}^\star = (\mu_t^\star)_{0\le t\le 1}$ with densities $\rho_t^\star$, such that:

1. $\mu_0^\star = \delta_0$ and $\mu_1^\star = \mu^\star$
2. The evolution of $t \mapsto \mu_t^\star$ is governed by two coupled PDEs:

$$\partial_t \rho_t^\star(w) = \nabla_w \cdot \big(\rho_t^\star(w) \nabla_w V^\star(w,t)\big) + \frac{\tau}{2} \Delta_w \rho_t^\star(w) \tag{1a}$$

$$\partial_t V^\star(w,t) = -\frac{\tau}{2} \Delta_w V^\star(w,t) + \frac{1}{2} \|\nabla_w V^\star(w,t)\|_2^2 \tag{1b}$$

   on $\mathbb{R}^d \times [0,1]$, where (1a) is the forward (Fokker–Planck) equation with initial condition $\rho_0^\star(\cdot) = \delta(\cdot)$ and (1b) is the backward (Hamilton–Jacobi–Bellman) equation with terminal condition

$$V^\star(w,1) = \beta \left(R_0 + \int \tilde{f} \, \mathrm{d}\mu_1^\star + \Psi(w; \mu_1^\star) \right).$$

3. $\boldsymbol{\mu}^\star$ arises as a solution of a stochastic control problem on the space of measures.

The McKean–Vlasov Control Problem

- Controlled Itô SDE:

$$\mathrm{d}W_t^{\boldsymbol{u}} = u_t\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad W_0 \equiv 0;\ t \in [0,1]$$

  where the control $\boldsymbol{u} = (u_t)_{t\in[0,1]}$ is a progressively measurable drift

The McKean–Vlasov Control Problem

- Controlled Itô SDE:

$$\mathrm{d}W_t^{\boldsymbol{u}} = u_t\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad W_0 \equiv 0;\ t \in [0,1]$$

  where the control $\boldsymbol{u} = (u_t)_{t\in[0,1]}$ is a progressively measurable drift
- Flow of measures $\boldsymbol{\mu}^{\boldsymbol{u}} = (\mu_t^{\boldsymbol{u}})_{t\in[0,1]}$, where $\mu_t^{\boldsymbol{u}} := \mathrm{Law}(W_t^{\boldsymbol{u}})$

The McKean–Vlasov Control Problem

- Controlled Itô SDE:

$$\mathrm{d}W_t^{\boldsymbol{u}} = u_t\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad W_0 \equiv 0;\ t \in [0,1]$$

  where the control $\boldsymbol{u} = (u_t)_{t\in[0,1]}$ is a progressively measurable drift
- Flow of measures $\boldsymbol{\mu}^{\boldsymbol{u}} = (\mu_t^{\boldsymbol{u}})_{t\in[0,1]}$, where $\mu_t^{\boldsymbol{u}} := \mathrm{Law}(W_t^{\boldsymbol{u}})$

Optimal control problem: minimize the total cost

$$J_{\beta,\tau}(\boldsymbol{u}) := \mathbf{E}\Bigg[\underbrace{\frac{1}{2}\int_0^1 \|u_t\|_2^2\,\mathrm{d}t}_{\text{control cost}}\Bigg] + \underbrace{\frac{\beta}{2}R(\mu_1^{\boldsymbol{u}})}_{\text{terminal cost}}$$

The McKean–Vlasov Control Problem

- Controlled Itô SDE:

$$\mathrm{d}W_t^{\boldsymbol{u}} = u_t\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad W_0 \equiv 0;\ t \in [0,1]$$

  where the control $\boldsymbol{u} = (u_t)_{t\in[0,1]}$ is a progressively measurable drift
- Flow of measures $\boldsymbol{\mu}^{\boldsymbol{u}} = (\mu_t^{\boldsymbol{u}})_{t\in[0,1]}$, where $\mu_t^{\boldsymbol{u}} := \mathrm{Law}(W_t^{\boldsymbol{u}})$

Optimal control problem: minimize the total cost

$$J_{\beta,\tau}(\boldsymbol{u}) := \mathbf{E}\Bigg[\underbrace{\frac{1}{2}\int_0^1 \|u_t\|_2^2\,\mathrm{d}t}_{\text{control cost}}\Bigg] + \underbrace{\frac{\beta}{2}R(\mu_1^{\boldsymbol{u}})}_{\text{terminal cost}}$$

- Key point: for a control $\boldsymbol{u}$, the terminal cost depends *nonlinearly* on the probability law of $W_1^{\boldsymbol{u}}$ — this is an instance of a *McKean–Vlasov control problem* (Carmona and Delarue)

Transport Map Interpretation

- Each control $\boldsymbol{u}$ *transports* the Brownian path $B_{[0,1]} = (B_t)_{t\in[0,1]}$ to a random vector $W_1^{\boldsymbol{u}}$:

$$T^{\boldsymbol{u}} : B_{[0,1]} \longmapsto \int_0^1 u_t \,\mathrm{d}t + \sqrt{\tau} B_1$$

Transport Map Interpretation

- Each control $\boldsymbol{u}$ *transports* the Brownian path $B_{[0,1]} = (B_t)_{t\in[0,1]}$ to a random vector $W_1^{\boldsymbol{u}}$:

$$T^{\boldsymbol{u}} : B_{[0,1]} \longmapsto \int_0^1 u_t \,\mathrm{d}t + \sqrt{\tau} B_1$$

- Zero drift ($\boldsymbol{u} \equiv 0$): $T^0(B_{[0,1]}) = \sqrt{\tau} B_1 \sim \gamma_\tau$ (samples from the Gaussian prior)

Transport Map Interpretation

- Each control $\boldsymbol{u}$ *transports* the Brownian path $B_{[0,1]} = (B_t)_{t\in[0,1]}$ to a random vector $W_1^{\boldsymbol{u}}$:

$$T^{\boldsymbol{u}} : B_{[0,1]} \longmapsto \int_0^1 u_t \,\mathrm{d}t + \sqrt{\tau} B_1$$

- Zero drift ($\boldsymbol{u} \equiv 0$): $T^0(B_{[0,1]}) = \sqrt{\tau} B_1 \sim \gamma_\tau$ (samples from the Gaussian prior)
- Any other $\boldsymbol{u}$ affects the distribution $\mu_1^{\boldsymbol{u}}$, incurs the control cost

$$\mathbf{E}\left[\frac{1}{2}\int_0^1 \|u_t\|_2^2 \,\mathrm{d}t\right] = \tau \cdot D_{\mathrm{KL}}(\boldsymbol{P}^{\boldsymbol{u}} \| \boldsymbol{P}^0),$$

where $\boldsymbol{P}^{\boldsymbol{u}} := \mathrm{Law}(W_{[0,1]}^{\boldsymbol{u}})$ and $\boldsymbol{P}^0 := \mathrm{Law}(B_{[0,1]})$

Transport Map Interpretation

- Each control $\boldsymbol{u}$ *transports* the Brownian path $B_{[0,1]} = (B_t)_{t\in[0,1]}$ to a random vector $W_1^{\boldsymbol{u}}$:

$$T^{\boldsymbol{u}} : B_{[0,1]} \longmapsto \int_0^1 u_t \,\mathrm{d}t + \sqrt{\tau} B_1$$

- Zero drift ($\boldsymbol{u} \equiv 0$): $T^0(B_{[0,1]}) = \sqrt{\tau} B_1 \sim \gamma_\tau$ (samples from the Gaussian prior)
- Any other $\boldsymbol{u}$ affects the distribution $\mu_1^{\boldsymbol{u}}$, incurs the control cost

$$\mathbf{E}\left[\frac{1}{2}\int_0^1 \|u_t\|_2^2 \,\mathrm{d}t\right] = \tau \cdot D_{\mathrm{KL}}(\boldsymbol{P}^{\boldsymbol{u}} \| \boldsymbol{P}^0),$$

  where $\boldsymbol{P}^{\boldsymbol{u}} := \mathrm{Law}(W^{\boldsymbol{u}}_{[0,1]})$ and $\boldsymbol{P}^0 := \mathrm{Law}(B_{[0,1]})$

- The *optimal control* $\boldsymbol{u}^\star$ will give $W_1^{\boldsymbol{u}^\star} \sim \mu^\star$, and

$$D_{\mathrm{KL}}(\boldsymbol{P}^{\boldsymbol{u}^\star} \| \boldsymbol{P}^0) = \min_{\boldsymbol{u}} D_{\mathrm{KL}}(\boldsymbol{P}^{\boldsymbol{u}} \| \boldsymbol{P}^0) = D_{\mathrm{KL}}(\mu^\star \| \gamma_\tau)$$

  — entropic optimal transport (or the Schrödinger bridge) problem!

Entropy-Regularized Risk: Dynamic Formulation

Theorem (Tzen–Raginsky, 2020) Let $\mu^\star$ be the (unique) minimizer of the free energy $F_{\beta,\tau}(\mu)$. Then the optimal controlled process solves the Itô SDE

$$\mathrm{d}W_t = -\nabla_w V^\star(W_t, t)\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad t \in [0,1];\ W_0 = 0,$$

where $V^\star(w,t) := -\tau \log \mathbf{E}\left[\exp\left(-\frac{\beta}{\tau}\Psi(B_\tau;\mu^\star)\right)\middle| B_{\tau t} = w\right]$.

Under the optimal control $\boldsymbol{u}^\star$, W_1 is distributed according to $\mu^\star$, and

$$\inf_{\boldsymbol{u}} J_{\beta,\tau}(\boldsymbol{u}) = \beta F_{\beta,\tau}(\mu^\star).$$

Entropy-Regularized Risk: Dynamic Formulation

Theorem (Tzen–Raginsky, 2020) Let $\mu^\star$ be the (unique) minimizer of the free energy $F_{\beta,\tau}(\mu)$. Then the optimal controlled process solves the Itô SDE

$$\mathrm{d}W_t = -\nabla_w V^\star(W_t, t)\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad t \in [0,1];\ W_0 = 0,$$

where $V^\star(w,t) := -\tau \log \mathbf{E}\left[\exp\left(-\frac{\beta}{\tau}\Psi(B_\tau;\mu^\star)\right)\middle| B_{\tau t} = w\right]$.

Under the optimal control $\boldsymbol{u}^\star$, W_1 is distributed according to $\mu^\star$, and

$$\inf_{\boldsymbol{u}} J_{\beta,\tau}(\boldsymbol{u}) = \beta F_{\beta,\tau}(\mu^\star).$$

The Schrödinger bridge problem:

- Let $\mathscr{U}(\mu^\star) := \{\boldsymbol{u} : W_1^{\boldsymbol{u}} \sim \mu^\star\}$
- Then the optimal drift $\boldsymbol{u}^\star \in \mathscr{U}(\mu^\star)$, and

$$\mathbf{E}\left[\frac{1}{2}\int_0^1 \|u_t^\star\|_2^2\,\mathrm{d}t\right] = \inf_{\boldsymbol{u}\in\mathscr{U}(\mu^\star)} \mathbf{E}\left[\frac{1}{2}\int_0^1 \|u_t\|_2^2\,\mathrm{d}t\right] = \tau D_{\mathrm{KL}}(\mu^\star\|\gamma_\tau)$$

A Closer Look at the Optimal Control

- Optimal McKean–Vlasov dynamics:

$$\mathrm{d}W_t = -\nabla_w V^\star(W_t, t)\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad t \in [0,1];\ W_0 = 0,$$

$$\text{where } V^\star(w,t) := -\tau \log \mathbf{E}\left[\exp\left(-\frac{\beta}{\tau} \Psi(B_\tau; \mu^\star) \right) \middle| B_{\tau t} = w \right].$$

A Closer Look at the Optimal Control

- Optimal McKean–Vlasov dynamics:

$$\mathrm{d}W_t = -\nabla_w V^\star(W_t, t)\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad t \in [0,1];\ W_0 = 0,$$

$$\text{where } V^\star(w,t) := -\tau \log \mathbf{E}\left[\exp\left(-\frac{\beta}{\tau}\Psi(B_\tau;\mu^\star)\right)\middle| B_{\tau t} = w\right].$$

- Direct computation:

$$-\nabla_w V^\star(W_t,t) = -\beta \int_{\mathbb{R}^d} \nabla\Psi(w;\mu^\star) Q_{W_t,t}(\mathrm{d}w)$$

where $Q_{w,t}(\mathrm{d}v)$ is the Gibbs measure

$$Q_{w,t}(A) := \frac{\int_A \exp\left(-\frac{\|v-w\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau}\Psi(v;\mu^\star)\right)\mathrm{d}v}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|v-w\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau}\Psi(v;\mu^\star)\right)\mathrm{d}v}$$

A Closer Look at the Optimal Control

- Optimal McKean–Vlasov dynamics:

$$\mathrm{d}W_t = -\nabla_w V^\star(W_t, t)\,\mathrm{d}t + \sqrt{\tau}\,\mathrm{d}B_t, \qquad t \in [0,1];\ W_0 = 0,$$

$$\text{where } V^\star(w,t) := -\tau \log \mathbf{E}\left[\exp\left(-\frac{\beta}{\tau}\Psi(B_\tau;\mu^\star)\right)\middle| B_{\tau t} = w\right].$$

- Direct computation:

$$-\nabla_w V^\star(W_t,t) = -\beta \int_{\mathbb{R}^d} \nabla\Psi(w;\mu^\star) Q_{W_t,t}(\mathrm{d}w)$$

where $Q_{w,t}(\mathrm{d}v)$ is the Gibbs measure

$$Q_{w,t}(A) := \frac{\int_A \exp\left(-\frac{\|v-w\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau}\Psi(v;\mu^\star)\right)\mathrm{d}v}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|v-w\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau}\Psi(v;\mu^\star)\right)\mathrm{d}v}$$

- Gibbs flow $(Q_{w,t})_{t\in[0,1]}$: interpolates between $Q_{0,0} = \mu^\star$ and $Q_{0,1} = \delta_0$; becomes more concentrated as $t \to 1$:

$$\int_{\mathbb{R}^d} \|v-w\|^2 Q_{w,t}(\mathrm{d}v) \le \kappa(\beta^2 + \tau d)(1-t).$$

Gradient Descent as a Greedy Heuristic

- Optimal McKean–Vlasov dynamics:

$$\mathrm{d}W_t = -\beta \left(\int_{\mathbb{R}^d} \nabla \Psi(w; \mu^\star) Q_{W_t,t}(\mathrm{d}w) \right) \mathrm{d}t + \sqrt{\tau}\, \mathrm{d}B_t$$

$$\text{where } Q_{W_t,t}(\mathrm{d}w) = \frac{\exp\left(-\frac{\|w - W_t\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau} \Psi(w; \mu^\star) \right) \mathrm{d}w}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|w - W_t\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau} \Psi(W_t; \mu^\star) \right) \mathrm{d}w}$$

  is concentrated around W_t

Gradient Descent as a Greedy Heuristic

- Optimal McKean–Vlasov dynamics:

$$\mathrm{d}W_t = -\beta \left(\int_{\mathbb{R}^d} \nabla \Psi(w; \mu^\star) Q_{W_t,t}(\mathrm{d}w) \right) \mathrm{d}t + \sqrt{\tau}\, \mathrm{d}B_t$$

$$\text{where } Q_{W_t,t}(\mathrm{d}w) = \frac{\exp\left(-\frac{\|w - W_t\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau} \Psi(w; \mu^\star) \right) \mathrm{d}w}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|w - W_t\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau} \Psi(W_t; \mu^\star) \right) \mathrm{d}w}$$

  is concentrated around W_t

- Greedy approximation:

$$\int_{\mathbb{R}^d} \nabla \Psi(w; \mu^\star) Q_{W_t,t}(\mathrm{d}w) \approx \nabla \Psi(W_t; \mu^\star) \approx \nabla \Psi(W_t; \mathrm{Law}(W_t))$$

Gradient Descent as a Greedy Heuristic

- Optimal McKean–Vlasov dynamics:

$$\mathrm{d}W_t = -\beta \left(\int_{\mathbb{R}^d} \nabla \Psi(w; \mu^\star) Q_{W_t,t}(\mathrm{d}w) \right) \mathrm{d}t + \sqrt{\tau}\, \mathrm{d}B_t$$

$$\text{where } Q_{W_t,t}(\mathrm{d}w) = \frac{\exp\left(-\frac{\|w - W_t\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau} \Psi(w; \mu^\star) \right) \mathrm{d}w}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|w - W_t\|_2^2}{2\tau(1-t)} - \frac{\beta}{\tau} \Psi(W_t; \mu^\star) \right) \mathrm{d}w}$$

  is concentrated around W_t

- Greedy approximation:

$$\int_{\mathbb{R}^d} \nabla \Psi(w; \mu^\star) Q_{W_t,t}(\mathrm{d}w) \approx \nabla \Psi(W_t; \mu^\star) \approx \nabla \Psi(W_t; \mathrm{Law}(W_t))$$

- Nonlinear dynamics:

$$\mathrm{d}\hat{W}_t = -\beta \nabla \Psi(\hat{W}_t; \mathrm{Law}(\hat{W}_t))\, \mathrm{d}t + \sqrt{\tau}\, \mathrm{d}B_t, \qquad \hat{W}_0 = 0;\ t \in [0, 1]$$

Question: can we show that $\hat{W}$ tracks W with high probability?

SGD as a Greedy Heuristic

SGD update:

- Data: $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} \pi, \quad Y_k = f(X_k)$
- Generate iterates $\boldsymbol{W}_k = (W_k^1, \ldots, W_k^N), \quad k = 0, 1, 2, \ldots$:

$$W_{k+1}^i = W_k^i + \eta\beta(Y_{k+1} - \hat{f}_N(X_{k+1}, \boldsymbol{W}_k))\nabla_w \sigma(X_{k+1}; W_k^i)$$

with $\boldsymbol{W}_0 \sim (\gamma_\tau)^{\otimes N}$ (Gaussian initialization)

SGD as a Greedy Heuristic

SGD update:

- Data: $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} \pi, \quad Y_k = f(X_k)$
- Generate iterates $\boldsymbol{W}_k = (W_k^1, \ldots, W_k^N), \quad k = 0, 1, 2, \ldots$:

$$W_{k+1}^i = W_k^i + \eta\beta(Y_{k+1} - \hat{f}_N(X_{k+1}, \boldsymbol{W}_k))\nabla_w \sigma(X_{k+1}; W_k^i)$$

with $\boldsymbol{W}_0 \sim (\gamma_\tau)^{\otimes N}$ (Gaussian initialization)

Note:

- no explicit regularization or additional noise
- Gaussian initialization, $W_0^i \overset{\text{i.i.d.}}{\sim} \gamma_\tau$
- we will be interested in the regime

$$\tau = \varepsilon^2/d, \qquad \beta = \sqrt{\tau d} = \varepsilon$$

SGD Tracks the Optimal McKean–Vlasov Dynamics

Theorem (Tzen–Raginsky, 2020) Let $\boldsymbol{\mu}^\star = \{\mu_t^\star\}_{t\in[0,1]}$ be the flow of measures along the optimal McKean–Vlasov dynamics, with $\mu_0^\star = \delta_0$ and $\mu_1^\star = \mu^\star$. Then, with probability at least $1-\delta$,

$$\max_{0\le k\le n} |R_N(\boldsymbol{W}_k) - R(\mu_{k\eta}^\star)| \le \frac{\kappa}{N} + \kappa\sqrt{\frac{1}{N}\log\frac{N}{\delta}} + \kappa\beta$$
$$+ \kappa(1+\sqrt{\eta}\beta)(1+\beta)e^{\kappa\beta}\left(\beta + \sqrt{\tau d \log\frac{Nd}{\delta\eta}}\right).$$

SGD Tracks the Optimal McKean–Vlasov Dynamics

Theorem (Tzen–Raginsky, 2020) Let $\boldsymbol{\mu}^\star = \{\mu_t^\star\}_{t\in[0,1]}$ be the flow of measures along the optimal McKean–Vlasov dynamics, with $\mu_0^\star = \delta_0$ and $\mu_1^\star = \mu^\star$. Then, with probability at least $1-\delta$,

$$\max_{0\le k\le n} |R_N(\boldsymbol{W}_k) - R(\mu_{k\eta}^\star)| \le \frac{\kappa}{N} + \kappa\sqrt{\frac{1}{N}\log\frac{N}{\delta}} + \kappa\beta$$
$$+ \kappa(1+\sqrt{\eta}\beta)(1+\beta)e^{\kappa\beta}\left(\beta + \sqrt{\tau d \log\frac{Nd}{\delta\eta}}\right).$$

Remarks:

- Similar to the bound of Mei–Misiakiewicz–Montanari (2019), except here $T = 1$
- If we take
$$\beta = \sqrt{\tau d} = \varepsilon \qquad \text{and} \qquad N \sim \frac{1}{\varepsilon^2},$$
then w.h.p. SGD will track optimal MKV dynamics to $O(\varepsilon)$ accuracy

Comparison with Nonlinear Dynamics

1. This work: Optimal McKean–Vlasov dynamics for $\mu_t^\star(\mathrm{d}w) = \rho_t^\star(w)\,\mathrm{d}w$

$$\partial_t \rho_t^\star(w) = \nabla_w \cdot \big(\rho_t^\star(w) \nabla_w V^\star(w,t)\big) + \frac{\tau}{2}\Delta_w \rho_t^\star(w)$$

$$\partial_t V^\star(w,t) = -\frac{\tau}{2}\Delta_w V^\star(w,t) + \frac{1}{2}\|\nabla_w V^\star(w,t)\|_2^2$$

(with appropriate initial and terminal conditions)

- finite time horizon ($T = 1$)
- $\mu_1^\star = \mu^\star$ (exact optimality at $T = 1$)
- two coupled nonlinear PDEs, forward and backward in time

Comparison with Nonlinear Dynamics

1. This work: Optimal McKean–Vlasov dynamics for $\mu_t^\star(\mathrm{d}w) = \rho_t^\star(w)\,\mathrm{d}w$

$$\partial_t \rho_t^\star(w) = \nabla_w \cdot \big(\rho_t^\star(w) \nabla_w V^\star(w,t)\big) + \frac{\tau}{2}\Delta_w \rho_t^\star(w)$$

$$\partial_t V^\star(w,t) = -\frac{\tau}{2}\Delta_w V^\star(w,t) + \frac{1}{2}\|\nabla_w V^\star(w,t)\|_2^2$$

   (with appropriate initial and terminal conditions)
   - finite time horizon ($T = 1$)
   - $\mu_1^\star = \mu^\star$ (exact optimality at $T = 1$)
   - two coupled nonlinear PDEs, forward and backward in time

2. Existing work: Nonlinear dynamics for $\mu_t(\mathrm{d}w) = \rho_t(w)\,\mathrm{d}w$

$$\partial_t \rho_t(w) = \beta \nabla_w \cdot \big(\rho_t(w) \nabla_w \Psi(w;\mu_t)\big) + \frac{\tau}{2}\Delta_w \rho_t(w)$$

   - infinite time ($T \to \infty$)
   - can guarantee convergence to $\mu^\star$ as $t \to \infty$
   - one PDE running forward in time

Comparison with Nonlinear Dynamics

1. This work: Optimal McKean–Vlasov dynamics for $\mu_t^\star(\mathrm{d}w) = \rho_t^\star(w)\,\mathrm{d}w$

$$\partial_t \rho_t^\star(w) = \nabla_w \cdot \big(\rho_t^\star(w) \nabla_w V^\star(w,t)\big) + \frac{\tau}{2}\Delta_w \rho_t^\star(w)$$

$$\partial_t V^\star(w,t) = -\frac{\tau}{2}\Delta_w V^\star(w,t) + \frac{1}{2}\|\nabla_w V^\star(w,t)\|_2^2$$

   (with appropriate initial and terminal conditions)
   - finite time horizon ($T = 1$)
   - $\mu_1^\star = \mu^\star$ (exact optimality at $T = 1$)
   - two coupled nonlinear PDEs, forward and backward in time

2. Existing work: Nonlinear dynamics for $\mu_t(\mathrm{d}w) = \rho_t(w)\,\mathrm{d}w$

$$\partial_t \rho_t(w) = \beta \nabla_w \cdot \big(\rho_t(w) \nabla_w \Psi(w;\mu_t)\big) + \frac{\tau}{2}\Delta_w \rho_t(w)$$

   - infinite time ($T \to \infty$)
   - can guarantee convergence to $\mu^\star$ as $t \to \infty$
   - one PDE running forward in time

With proper scaling of τ and β, nonlinear dynamics can be seen as a greedy approximation to optimal MKV dynamics.

Preprint at https://arxiv.org/abs/2002.01987