Constrained sampling via Langevin dynamics

Volkan Cevher
https://lions.epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

École Polytechnique, Paris

[June 2019]

Statistical Inference: A summary-based approach

- M-estimators in statistics/learning: Assume prior $p(x)$, observe data $z_1, z_2, ..., z_n$.
  - ▷ Find the best parameter/predictor via *optimization*:

$$\hat{x} = \arg\min_x \sum_{i=1}^{n} f(x; z_i) - \log p(x).$$

  - ▷ An intuitive summary of the posterior
  - ▷ Extremely popular (ex., MLE, ERM, etc.) with theoretical backup

Statistical Inference: A more descriptive approach

- *Posterior* sampling: Assume prior $p(x)$, observe data $z_1, z_2, ..., z_n$.
  - ▷ Generate samples from the *posterior*

$$p(x|z_1, z_2, ..., z_n) \propto p(x) \cdot \exp\left(-\sum_{i=1}^{n} f(x; z_i)\right).$$

  - ▷ Ability to obtain different summaries of the posterior (i.e., posterior mean)
  - ▷ Reinventing itself

Point estimates vs posterior sampling

- Posterior sampling: Arguably more difficult.
- Upshots
  - ▷ Flexibility: Generativeness (i.e., Generative Adversarial Networks)...
  - ▷ Uncertainty: Distribution of parameters/predictors, confidence bounds...
  - ▷ Information: Multi-modality, other summaries and features...

Summary of this lecture

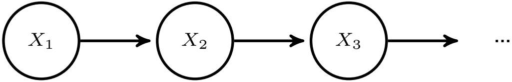
- Unconstrained sampling via unadjusted Langevin Dynamics
- Constrained sampling via mirrored Langevin dynamics
- Applications / Extensions:
    - ▷ Sampling for Generative Adversarial Networks
    - ▷ Sampling for non-convex optimization

Before we begin, let's recall a key concept

- Markov chain [Kemeny and Snell, 1976]
  - ▷ sequence of random variables $\{X_1, X_2, \ldots\}$ satisfying

$$\mathsf{Pr}(X_i | X_1, X_2, \ldots, X_{i-1}) = \mathsf{Pr}(X_i | X_{i-1}) \quad \forall i > 1$$

- Example Bayesian network



  - ▷ remove arrows to form a Markov random field

- No memory: Future is independent of the past given the present.

Markov Chain Monte Carlo

- Goal: Sample from a distribution μ.
- Idea of MCMC:
  - ▷ Construct a Markov Chain whose stationary distribution is μ
  - ▷ Simulate the Markov Chain until mixing time
- Key question: *How to design such a Markov Chain?*

Markov Chain Monte Carlo

- Goal: Sample from a distribution μ.
- Idea of MCMC:
  - ▷ Construct a Markov Chain whose stationary distribution is μ
  - ▷ Simulate the Markov Chain until mixing time
- Key question: *How to design such a Markov Chain?*
- Many possibilities:
  - ▷ Metropolis-Hasting [Metropolis and Ulam, 1949, Hastings, 1970]
  - ▷ Gibbs sampling [Gelfand et al., 1990]
  - ▷ Hamiltonian Monte Carlo [Neal et al., 2011]
  - ▷ Importance sampling [Kahn and Harris, 1951]
  - ▷ ...

Langevin Monte Carlo

- Task: Given a target distribution $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$, generate samples from μ.
  - ▷ Most samples should be gathered around the minimum of V
  - ▷ We do not want convergence to the minimum

- Idea: Use Gradient Descent + noise

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \beta_k \nabla V(\mathbf{X}^k) + \text{noise}$$

- Question: What amount of noise should we add so that $\mathbf{X}^\infty \sim \pi$?

Unadjusted Langevin Algorithm

- Task: Given a target distribution $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$, generate samples from μ.
  - ▷ Fundamental in machine learning/statistics/computer science/etc.

- A scalable framework: First-order sampling (assuming access to ∇V).

  Step 1. **Langevin Dynamics**

$$\mathrm{d}\mathbf{X}_t = -\nabla V(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t \quad \Rightarrow \quad \mathbf{X}_\infty \sim e^{-V}.$$

  Step 2. **Discretize**

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \beta_k \nabla V(\mathbf{x}^k) + \sqrt{2\beta_k}\boldsymbol{\xi}^k$$

  - ▷ β_k step-size, $\boldsymbol{\xi}^k$ standard normal
  - ▷ strong analogy to gradient descent method

Log-concave distribution

Definition

A function $V : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{X}$ if, for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{X}$ and $\alpha \in [0, 1]$, we have:

$$V(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha V(\mathbf{x}_1) + (1-\alpha)V(\mathbf{x}_2).$$

Corollary

If $V : \mathcal{X} \to \mathbb{R}$ is twice differentiable, then V is convex if and only if

$$\nabla^2 V(\mathbf{x}) \succeq 0\ \forall \mathbf{x} \in \mathcal{X},$$

where $\nabla^2 V(x)$ denotes the Hessian matrix of V at $\mathbf{x}$, and $\forall A, B \in \mathbb{R}^{d\times d}$, $A \succeq B$ means that $A - B$ is positive semi-definite.

- Similarly, a function V is called concave if $-V$ is convex.

Definition

A distribution μ over $\mathcal{X}$ is called log-concave if it can be written as $\mathrm{d}\mu = e^{-V(\mathbf{x})}\mathrm{d}\mathbf{x}$ with V convex.

Recent progress: Unconstrained distributions are easy

- State-of-the-art: When $\mathrm{dom}(V) = \mathbb{R}^d$,

Assumption	$\mathcal{W}_2$	d_{TV}	KL	Literature
$LI \succeq \nabla^2 V \succeq mI$	$\tilde{O}\left(\epsilon^{-2} d\right)$	$\tilde{O}\left(\epsilon^{-2} d\right)$	$\tilde{O}\left(\epsilon^{-1} d\right)$	[Cheng and Bartlett, 2017] [Dalalyan and Karagulyan, 2017] [Durmus et al., 2018]
$LI \succeq \nabla^2 V \succeq 0$	-	$\tilde{O}\left(\epsilon^{-4} d\right)$	$\tilde{O}\left(\epsilon^{-2} d\right)$	[Durmus et al., 2018]

Note: $\mathcal{W}_2(\mu_1, \mu_2) \coloneqq \sqrt{\inf_{X\sim\mu_1, Y\sim\mu_2} \mathbb{E}\|X - Y\|^2}, \quad d_{\mathsf{TV}}(\mu_1, \mu_2) \coloneqq \sup_{A \text{ Borel}} |\mu_1(A) - \mu_2(A)|$

- Added twists with additional assumptions
  - ▷ Metropolis-Hastings [Dwivedi et al., 2018]
  - ▷ Underdamping [Cheng et al., 2017]
  - ▷ JKO-based schemes [Wibisono, 2018, Bernton, 2018, Jordan et al., 1998]

- What about **constrained** distributions?
  - ▷ include many important applications, such as Latent Dirichlet Allocation (LDA)

A challenge: Constrained distributions are hard

- When $\mathrm{dom}(V)$ is compact, convergence rates deteriorate significantly.

Assumption	$\mathcal{W}_2$ or KL	d_{TV}	Literature
$LI \succeq \nabla^2 V \succeq mI$	?	$\tilde{O}\left(\epsilon^{-6} d^5\right)$	[Brosse et al., 2017]
$LI \succeq \nabla^2 V \succeq 0$	?	$\tilde{O}\left(\epsilon^{-6} d^5\right)$	[Brosse et al., 2017]

  - ▷ *cf.*, when V is unconstrained, $\tilde{O}(\epsilon^{-4} d)$ convergence in d_{TV}
  - ▷ Projection is **not** a solution: Slow rates [Bubeck et al., 2015], boundary issues

Another challenge: Convergence for non-log-concave distributions

- Log-concave distributions have nice convergence.

- What about *non-log-concave* distributions?

  ▷ The **Dirichlet posteriors**:

$$V(x) = C - \sum_{\ell=1}^{d+1} (n_\ell + \alpha_\ell - 1) \log x_\ell.$$

  ▷ Super important in text modeling (via LDA), but no existing rate.

Mirrored Langevin Dynamics [Hsieh et al., 2018a]

- We provide a unified framework for **constrained** sampling.
  - ▷ Provide matching rates as if the constraint is absent: $\tilde{O}\left(\epsilon^{-4}d^{4}\right)$ improvement

- The first non-asymptotic guarantee for Dirichlet posteriors (in first-order sampling).
  - ▷ Applicable to a wide class of distributions

- Mini-batch or on-line extensions: sampling with stochastic gradients.
  - ▷ Experiments for the **Latent Dirichlet Allocation** (LDA), Wikipedia corpus

approach, algorithm, and theory

A starting point: [Nemirovski-Yudin 83] + [Beck-Teboulle 03]

- Entropic Mirror Descent: Unconstrained optimization within the simplex

$$x^{k+1} = \nabla h^* \left(\nabla h(x^k) - \beta_k \nabla V(x^k)\right), \quad (h: \text{ the entropy function})$$

▷ Mirror vs primal image: $y = \nabla h(x) \Leftrightarrow x = \nabla h^*(y)$

$$y^{k+1} = y^k - \beta_k \nabla V(x^k) \Rightarrow \text{ no projection in the mirrored space}$$

▷ Compare to the projected gradient algorithm

$$x^{k+1} = \mathsf{Proj}_\Delta \left(x^k - \beta_k \nabla V(x^k)\right).$$

A starting point: [Nemirovski-Yudin 83] + [Beck-Teboulle 03]

- Entropic Mirror Descent: Unconstrained optimization within the simplex

$$x^{k+1} = \nabla h^* \left(\nabla h(x^k) - \beta_k \nabla V(x^k) \right), \quad (h: \text{ the entropy function})$$

▷ Mirror vs primal image: $y = \nabla h(x) \Leftrightarrow x = \nabla h^*(y)$

$$y^{k+1} = y^k - \beta_k \nabla V(x^k) \Rightarrow \text{ no projection in the mirrored space}$$

▷ Compare to the projected gradient algorithm

$$x^{k+1} = \mathrm{Proj}_{\Delta} \left(x^k - \beta_k \nabla V(x^k) \right).$$

Q: Can we derive an entropic, **unconstrained** *Langevin dynamics on the simplex?*

A starting point: [Nemirovski-Yudin 83] + [Beck-Teboulle 03]

- Entropic Mirror Descent: Unconstrained optimization within the simplex

$$x^{k+1} = \nabla h^* \left(\nabla h(x^k) - \beta_k \nabla V(x^k)\right), \quad (h: \text{ the entropy function})$$

▷ Mirror vs primal image: $y = \nabla h(x) \Leftrightarrow x = \nabla h^*(y)$

$$y^{k+1} = y^k - \beta_k \nabla V(x^k) \Rightarrow \text{ no projection in the mirrored space}$$

▷ Compare to the projected gradient algorithm

$$x^{k+1} = \mathsf{Proj}_{\Delta} \left(x^k - \beta_k \nabla V(x^k)\right).$$

Q: Can we derive an entropic, **unconstrained** *Langevin dynamics on the simplex?*

- More generally, a "mirror descent theory" for Langevin Dynamics? Any benefit?

Mirrored Langevin Dynamics MLD

- Approach: Push-forward measure

$$\nu = \nabla h \# \mu \Leftrightarrow \nu(A) = \mu(\nabla h^{-1}(A)),\ A \text{ Borel.}$$

- A simple but crucial observation: Let $Y_t = \nabla h(X_t)$, then it holds

$$h \text{ strictly convex } \Leftrightarrow \nabla h \text{ one-to-one } \Rightarrow Y_\infty = \nabla h(X_\infty) \sim \nabla h \# e^{-V}.$$

Mirrored Langevin Dynamics MLD

- Approach: Push-forward measure

$$\nu = \nabla h \# \mu \Leftrightarrow \nu(A) = \mu(\nabla h^{-1}(A)), \; A \text{ Borel}.$$

- A simple but crucial observation: Let $Y_t = \nabla h(X_t)$, then it holds

$$h \text{ strictly convex } \Leftrightarrow \nabla h \text{ one-to-one } \Rightarrow Y_\infty = \nabla h(X_\infty) \sim \nabla h \# e^{-V}.$$

- For ease of discretization, simply consider the **Langevin Dynamics** for $\nabla h \# e^{-V}$:

$$\text{MLD} \equiv \begin{cases} dY_t = -\nabla W \bullet \nabla h(X_t)dt + \sqrt{2}dB_t \\ X_t = \nabla h^*(Y_t) \end{cases}, \text{ where } e^{-W} = \nabla h \# e^{-V}.$$

Implications of MLD I: Preserving the convergence

- Theory: Sampling with or without constraint has the same iteration complexity.

Theorem (Informal)

For discretized MLD with strongly convex h,

1. *If y^k is close to e^{-W} in KL/$\mathcal{W}_1/\mathcal{W}_2/d_{\mathsf{TV}}$, then x^k is close to e^{-V} in KL/$\mathcal{W}_1/\mathcal{W}_2/d_{\mathsf{TV}}$.*
2. *Let $\nabla^2 V \succeq mI$, bounded convex domain. Then there exists a good mirror map such that the discretized MLD yields convergence*

$$\tilde{O}\left(\epsilon^{-1}d\right) \text{ in KL}, \quad \tilde{O}\left(\epsilon^{-2}d\right) \text{ in } d_{\mathsf{TV}}, \quad \tilde{O}\left(\epsilon^{-2}d\right) \text{ in } \mathcal{W}_1, \quad \tilde{O}\left(\epsilon^{-2}d\right) \text{ in } \mathcal{W}_2.$$

- Comparison to state-of-the-art:

Assumption	$\mathcal{W}_1$	d_{TV}	Literature
$LI \succeq \nabla^2 V \succeq mI$	$\tilde{O}\left(\epsilon^{-6}d^5\right)$	$\tilde{O}\left(\epsilon^{-6}d^5\right)$	[Brosse et al., 2017]

Implications of MLD I: Preserving the convergence

- Theory: Sampling with or without constraint has the same iteration complexity.

Theorem (Informal)

For discretized MLD with strongly convex h,

1. *If y^k is close to e^{-W} in KL/$\mathcal{W}_1$/$\mathcal{W}_2$/d_{TV}, then x^k is close to e^{-V} in KL/$\mathcal{W}_1$/$\mathcal{W}_2$/d_{TV}.*
2. *Let $\nabla^2 V \succeq mI$, bounded convex domain. Then there exists a good mirror map such that the discretized MLD yields convergence*

$$\tilde{O}\left(\epsilon^{-1}d\right) \text{ in KL}, \quad \tilde{O}\left(\epsilon^{-2}d\right) \text{ in } d_{\mathsf{TV}}, \quad \tilde{O}\left(\epsilon^{-2}d\right) \text{ in } \mathcal{W}_1, \quad \tilde{O}\left(\epsilon^{-2}d\right) \text{ in } \mathcal{W}_2.$$

- Comparison to state-of-the-art:

Assumption	$\mathcal{W}_1$	d_{TV}	Literature
$LI \succeq \nabla^2 V \succeq mI$	$\tilde{O}\left(\epsilon^{-6}d^5\right)$	$\tilde{O}\left(\epsilon^{-6}d^5\right)$	[Brosse et al., 2017]

Caveat: Existential; no explicit algorithm.

Implications of MLD I: Preserving the convergence

- A key lemma: Let $B_h(x, x') = h(x) - h(x') - \langle \nabla h(x'), x - x' \rangle$.

Lemma (Duality of Wasserstein Distances)

Let μ_1, μ_2 be probability measures. Then, it holds that

$$\mathcal{W}_{B_h}(\mu_1, \mu_2) = \mathcal{W}_{B_{h^*}}(\nabla h \# \mu_1, \nabla h \# \mu_2),$$

where

$$\mathcal{W}_{B_h}(\mu_1, \mu_2) \coloneqq \inf_{T: T\#\mu_1 = \mu_2} \int B_h(x, T(x)) d\mu_1(x),$$

$$\mathcal{W}_{B_{h^*}}(\nu_1, \nu_2) \coloneqq \inf_{T: T\#\nu_1 = \nu_2} \int B_{h^*}(T(y), y) d\nu_1(y).$$

- Generalizes the classical duality:

$$B_h(x, x') = B_{h^*}(\nabla h(x'), \nabla h(x))$$

Convergence in $\mathcal{W}_2$

Theorem (Formal, $\mathcal{W}_2$)

For discretized MLD with ρ-strongly convex h,

$$\mathcal{W}_2(x^t, e^{-V}) \leq \frac{1}{\rho}\mathcal{W}_2(y^t, e^{-W}).$$

Convergence in $\mathcal{W}_2$

Theorem (Formal, $\mathcal{W}_2$)

For discretized MLD with ρ-strongly convex h,

$$\mathcal{W}_2(x^t, e^{-V}) \leq \frac{1}{\rho}\mathcal{W}_2(y^t, e^{-W}).$$

- Classical mirror descent theory:
  - If h is ρ-strongly convex, then

$$\frac{\rho}{2}\|x - x'\|^2 \leq B_h(x, x') = B_{h^*}(\nabla h(x'), \nabla h(x)) \leq \frac{1}{2\rho}\|\nabla h(x) - \nabla h(x')\|^2.$$

  - Take expectation, use $y^t \sim \nabla h \# x^t$ and $e^{-W} = \nabla h \# e^{-V}$
  - Our key lemma asserts the $=$ holds in the Wasserstein space. $\square$

Implications of MLD II: Practical sampling on simplex

- Practice: Efficient sampling algorithms on the simplex
  - ▷ Run A-MLD with the entropic mirror map

$$h(x) = \sum_{i=1}^{d} x_i \log x_i + \left(1 - \sum_{i=1}^{d} x_i\right) \log \left(1 - \sum_{i=1}^{d} x_i\right)$$

Implications of MLD II: Practical sampling on simplex

- Practice: Efficient sampling algorithms on the simplex
  - ▷ Run A-MLD with the entropic mirror map

$$h(x) = \sum_{i=1}^{d} x_i \log x_i + \left(1 - \sum_{i=1}^{d} x_i\right) \log\left(1 - \sum_{i=1}^{d} x_i\right)$$

  - ▷ Explicit formula for $\nabla h \# e^{-V}$:

$$W(y) = V \circ \nabla h^*(y) - \sum_{i=1}^{d} y_i + (d+1) h^*(y), \quad h^*(y) = \log\left(1 + \sum_{i=1}^{d} e^{y_i}\right).$$

Implications of MLD II: Practical sampling on simplex

- Practice: Efficient sampling algorithms on the simplex
  - ▷ Run A-MLD with the entropic mirror map

$$h(x) = \sum_{i=1}^{d} x_i \log x_i + \left(1 - \sum_{i=1}^{d} x_i\right) \log \left(1 - \sum_{i=1}^{d} x_i\right)$$

  - ▷ Explicit formula for $\nabla h \# e^{-V}$:

$$W(y) = V \circ \nabla h^*(y) - \sum_{i=1}^{d} y_i + (d+1) h^*(y), \quad h^*(y) = \log \left(1 + \sum_{i=1}^{d} e^{y_i}\right).$$

  - ▷ Surprise: Convex W vs. non-convex V!

The curious case of the Dirichlet Posterior

- Dirichlet posterior: $V(x) = C - \sum_{i=1}^{d+1} (n_i + \alpha_i - 1) \log x_i$
  - ▷ n_i (observations), α_i (parameters)
  - ▷ A simple approach in topic modeling
  - ▷ A practical (sparse) scenario: $n_1 = 10000, n_2 = 10, n_3 = n_4 = \cdots = 0,$

    $$\alpha_i = 0.1 \text{ for all } i$$

  - ▷ V is neither convex nor concave $\Rightarrow$ existing theory does not apply.

The curious case of the Dirichlet Posterior

- Dirichlet posterior: $V(x) = C - \sum_{i=1}^{d+1}(n_i + \alpha_i - 1)\log x_i$
  - ▷ n_i (observations), α_i (parameters)
  - ▷ A simple approach in topic modeling
  - ▷ A practical (sparse) scenario: $n_1 = 10000, n_2 = 10, n_3 = n_4 = \cdots = 0,$

$$\alpha_i = 0.1 \text{ for all } i$$

  - ▷ V is neither convex nor concave $\Rightarrow$ existing theory does not apply.

- For MLD,

$$W(y) = C' - \sum_{i=1}^{d}(n_i + \alpha_i)y_i + \left(\sum_{i=1}^{d+1}(n_i + \alpha_i)\right) h^*(y)$$

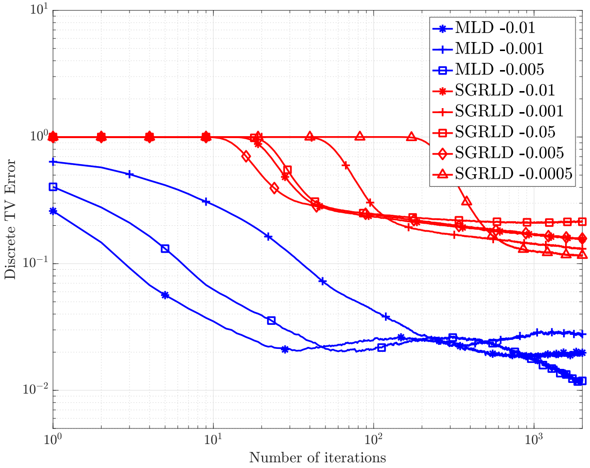
(i) W convex and Lipschitz gradient

  - ▷ $\tilde{O}(\epsilon^{-2}d^2R_0)$ convergence in KL, for any setting of n_i's and α_i's!

(ii) **Unconstrained** W: Efficient algorithm without projections

MLD for Dirichlet Posterior

- Synthetic setup: 11-dimensional Dirichlet posterior on the simplex
  - ▷ Observations $n_1 = 10000$, $n_2 = n_3 = 10$, $n_4 = n_5 = \cdots n_{11} = 0$.
  - ▷ Parameters $\alpha_i = 0.1$ for all i
  - ▷ Comparing against SGRLD [Patterson and Teh, 2013]



Stochastic MLD

Stochastic first-order sampling

- Evaluating ∇V is expensive in contemporary ML

- Simple solution: Use stochastic gradient estimates $\tilde{\nabla} V$.
  - ▷ Already present in the seminal work of [Welling and Teh, 2011].
  - ▷ Wide empirical success.
  - ▷ Some theory for *unconstrained* log-concave distributions.

Stochastic Mirrored Langevin Dynamics

- Assume decomposibility: $V = \sum_{i=1}^{N} V_i$.
  - ▷ Stochastic gradient $\tilde{\nabla} V = \frac{N}{b} \nabla V_i$ for a random mini-batch B, $|B| = b$.

Stochastic Mirrored Langevin Dynamics

- Assume decomposibility: $V = \sum_{i=1}^{N} V_i$.
  - ▷ Stochastic gradient $\tilde{\nabla} V = \frac{N}{b} \nabla V_i$ for a random mini-batch B, $|B| = b$.

Q: An MLD with only stochastic gradients? Convergence?

Stochastic Mirrored Langevin Dynamics

- Assume decomposibility: $V = \sum_{i=1}^{N} V_i$.
  - ▷ Stochastic gradient $\tilde{\nabla} V = \frac{N}{b} \nabla V_i$ for a random mini-batch B, $|B| = b$.

Lemma (SMLD)

Assume that h is 1-strongly convex. For i $= 1, 2, ..., N$, let W_i be such that

$$e^{-NW_i} = \nabla h \# \frac{e^{-NV_i}}{\int e^{-NV_i}}.$$

Define $W := \sum_{i=1}^{N} W_i$ and $\tilde{\nabla} W := \frac{N}{b} \sum_{i \in B} \nabla W_i$. Then:

1. *Primal decomposibility implies dual decomposability: There is a constant C such that $e^{-(W+C)} = \nabla h \# e^{-V}$.*

Stochastic Mirrored Langevin Dynamics

- Assume decomposibility: $V = \sum_{i=1}^{N} V_i$.
  - ▷ Stochastic gradient $\tilde{\nabla} V = \frac{N}{b} \nabla V_i$ for a random mini-batch B, $|B| = b$.

Lemma (SMLD)

Assume that h is 1-strongly convex. For i $= 1, 2, ..., N$, let W_i be such that

$$e^{-NW_i} = \nabla h \# \frac{e^{-NV_i}}{\int e^{-NV_i}}.$$

Define $W := \sum_{i=1}^{N} W_i$ and $\tilde{\nabla} W := \frac{N}{b} \sum_{i \in B} \nabla W_i$. Then:

1. *Primal decomposibility implies dual decomposability: There is a constant C such that $e^{-(W+C)} = \nabla h \# e^{-V}$.*
2. *For each i, the gradient ∇W_i depends only on ∇V_i and the mirror map h.*

Stochastic Mirrored Langevin Dynamics

- Assume decomposibility: $V = \sum_{i=1}^{N} V_i$.
  - ▷ Stochastic gradient $\tilde{\nabla} V = \frac{N}{b} \nabla V_i$ for a random mini-batch B, $|B| = b$.

Lemma (SMLD)

Assume that h is 1-strongly convex. For i $= 1, 2, ..., N$, let W_i be such that

$$e^{-NW_i} = \nabla h \# \frac{e^{-NV_i}}{\int e^{-NV_i}}.$$

Define $W := \sum_{i=1}^{N} W_i$ and $\tilde{\nabla} W := \frac{N}{b} \sum_{i \in B} \nabla W_i$. Then:

1. *Primal decomposibility implies dual decomposability: There is a constant C such that $e^{-(W+C)} = \nabla h \# e^{-V}$.*
2. *For each i, the gradient ∇W_i depends only on ∇V_i and the mirror map h.*
3. *The gradient estimate is unbiased: $\mathbb{E} \tilde{\nabla} W = \nabla W$.*

Stochastic Mirrored Langevin Dynamics

- Assume decomposibility: $V = \sum_{i=1}^{N} V_i$.
  - ▷ Stochastic gradient $\tilde{\nabla} V = \frac{N}{b} \nabla V_i$ for a random mini-batch B, $|B| = b$.

Lemma (SMLD)

Assume that h is 1-strongly convex. For i $= 1, 2, ..., N$, let W_i be such that

$$e^{-NW_i} = \nabla h \# \frac{e^{-NV_i}}{\int e^{-NV_i}}.$$

Define $W := \sum_{i=1}^{N} W_i$ and $\tilde{\nabla} W := \frac{N}{b} \sum_{i \in B} \nabla W_i$. Then:

1. *Primal decomposibility implies dual decomposability: There is a constant C such that $e^{-(W+C)} = \nabla h \# e^{-V}$.*
2. *For each i, the gradient ∇W_i depends only on ∇V_i and the mirror map h.*
3. *The gradient estimate is unbiased: $\mathbb{E} \tilde{\nabla} W = \nabla W$.*
4. *The dual stochastic gradient is never less accurate:*

$$\mathbb{E} \| \tilde{\nabla} W - \nabla W \|^2 \leq \mathbb{E} \| \tilde{\nabla} V - \nabla V \|^2.$$

Stochastic Mirrored Langevin Dynamics

Theorem (Convergence of SMLD)

Let e^{-V} *be the target distribution and* h *a 1-strongly convex mirror map. Let* $\sigma^2 := \mathbb{E}\|\tilde{\nabla}V - \nabla V\|^2$ *be the variance of the stochastic gradient of* V*. Suppose that* $LI \succeq \nabla^2 W \succeq 0$*. Then, applying SMLD with constant step-size* $\beta^t = \beta$ *yields:*

$$D(x^T \| e^{-V}) \leq \sqrt{\frac{2\mathcal{W}_2^2(y^0, e^{-W})(Ld + \sigma^2)}{T}} = O\left(\sqrt{\frac{Ld + \sigma^2}{T}}\right), \qquad (1)$$

provided that $\beta \leq \min\left\{\left[2T\mathcal{W}_2{}^2(\mathbf{y}^0, e^{-W})\left(Ld + \sigma^2\right)\right]^{-\frac{1}{2}}, \frac{1}{L}\right\}$.

Stochastic Mirrored Langevin Dynamics

Theorem (Convergence of SMLD)

Let e^{-V} be the target distribution and h a 1-strongly convex mirror map. Let $\sigma^2 := \mathbb{E}\|\tilde{\nabla} V - \nabla V\|^2$ be the variance of the stochastic gradient of V. Suppose that $LI \succeq \nabla^2 W \succeq 0$. Then, applying SMLD with constant step-size $\beta^t = \beta$ yields:

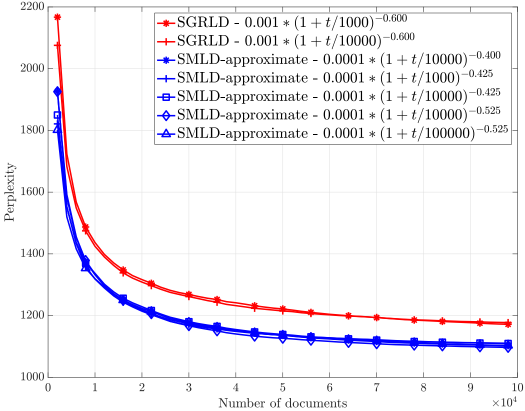
$$D(x^T \| e^{-V}) \leq \sqrt{\frac{2\mathcal{W}_2^2(y^0, e^{-W})(Ld + \sigma^2)}{T}} = O\left(\sqrt{\frac{Ld + \sigma^2}{T}}\right), \tag{1}$$

provided that $\beta \leq \min\left\{ \left[2T\mathcal{W}_2{}^2(\mathbf{y}^0, e^{-W}) \left(Ld + \sigma^2\right)\right]^{-\frac{1}{2}}, \frac{1}{L} \right\}$.

- For Dirichlet posteriors,
  - ▷ W is strictly convex.
  - ▷ $L = \sum_{\ell=1}^{d+1} n_\ell + \alpha_\ell = O(N + d)$.
  - ▷ simple formula for W_i's.
  - ▷ first non-asymptotic rate for first-order sampling.

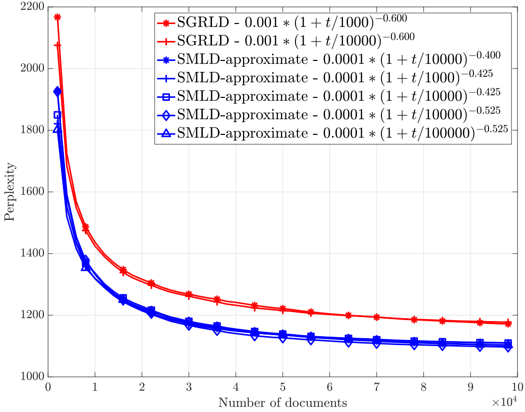
SMLD for Latent Dirichlet Allocation (LDA)

- LDA: A key framework in topic modeling.
  - ▷ Involves multiple Dirichlet posteriors (1 for each topic).
  - ▷ Wikipedia corpus with 100000 documents, 100 topics.



SMLD for Latent Dirichlet Allocation (LDA)

- LDA: A key framework in topic modeling.
  - ▷ Involves multiple Dirichlet posteriors (1 for each topic).
  - ▷ Wikipedia corpus with 100000 documents, 100 topics.



Caveat: Need to use $e^y \simeq \max\{0, 1+y\}$ in large-scale application.

Summary and follow-ups

- A new mirror descent theory for first-order sampling
  - ▷ Take any unconstrained algorithm for e^{-W} to improve efficiency

    Metropolis-Hastings adjusted Langevin Dynamics [Dwivedi et al., 2018]

    $\Rightarrow$ Linear rate for e^{-V} with $\nabla^2 V \succeq mI$, constrained or not.
  - ▷ Nontrivial extension of ideas & techniques from mirror descent
  - ▷ Stochastic version works well on real datasets

- We operate with the non-ideal ℓ_2-norm for entropic mirror map
  - ▷ Does not really concern practitioners, but unsatisfactory in theory.
  - ▷ Future work: Provide the ℓ_1 (or ℓ_p) counterpart.

- Need more theoretically guided step-size rules coming out of the theory
  - ▷ Working on it!

Application of SGLD to GANs training

Before we begin, let us review a key concept: Minimax problems

- Optimization template:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) \coloneqq f(x^*, y^*), \quad \mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} \subset \mathbb{R}^n.$$

- Sion's Minimax Theorem [Sion, 1958]: Under mild conditions,

$$f \text{ convex in } x \text{ and concave in } y \quad \Rightarrow \quad \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y).$$

Before we begin, let us review a key concept: Minimax problems

- Optimization template:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) \coloneqq f(x^*, y^*), \quad \mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} \subset \mathbb{R}^n.$$

- Sion's Minimax Theorem [Sion, 1958]: Under mild conditions,

$$f \text{ convex in } x \text{ and concave in } y \quad \Rightarrow \quad \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y).$$

- Numerous applications.
  - ▷ Langrange dual for constrained optimization
  - ▷ Robust optimization
  - ▷ Image denoising
  - ▷ Game theory
  - ▷ etc.

Saddle-Point Problems (cont.)

- Optimization template:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) \coloneqq f(x^*, y^*), \quad \mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} \subset \mathbb{R}^n.$$

- Algorithms for finding (x^*, y^*) when f convex in x and concave in y.
  - ▷ Prox methods: [Nemirovskii and Yudin, 1983, Nemirovski, 2004, Beck and Teboulle, 2003]
  - ▷ Chambolle-Pock [Chambolle and Pock, 2011]
  - ▷ Fast rates **assuming efficiently solvable iterates**.

Saddle-Point Problems (cont.)

- Optimization template:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) \coloneqq f(x^*, y^*), \quad \mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} \subset \mathbb{R}^n.$$

- Algorithms for finding (x^*, y^*) when f convex in x and concave in y.
  - ▷ Prox methods: [Nemirovskii and Yudin, 1983, Nemirovski, 2004, Beck and Teboulle, 2003]
  - ▷ Chambolle-Pock [Chambolle and Pock, 2011]
  - ▷ Fast rates **assuming efficiently solvable iterates**.

- In short, for such optimization,
  - ▷ The problem is well-defined (saddle-points exist).
  - ▷ Efficient algorithms with rigorous convergence rates.

Open problems: Non-convex/non-concave Saddle-Point Problems

- Optimization template:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) := f(x^*, y^*), \quad \mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} \subset \mathbb{R}^n.$$

- What if f is NOT convex in x and NOT concave in y?
  - ▷ Does a (local) saddle-point exist? Under what conditions??
  - ▷ Algorithms with convergence rates???

Main Motivation: Generative Adversarial Networks (GANs)

- Objective of Wasserstein GANs:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} [D_w(X)] .$$

  - ▷ θ: a **generator** neural net, generating fake data.
  - ▷ w: a **discriminator** neural net.
  - ▷ D_w: output of discriminator at w, *highly* non-convex/non-concave.

Main Motivation: Generative Adversarial Networks (GANs)

- Objective of Wasserstein GANs:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} [D_w(X)] .$$

  - ▷ θ: a **generator** neural net, generating fake data.
  - ▷ w: a **discriminator** neural net.
  - ▷ D_w: output of discriminator at w, *highly* non-convex/non-concave.

- Theoretical challenge:
  - ▷ A saddle point might NOT exist [Dasgupta and Maskin, 1986] (non-existence of pure strategy equilibrium).
  - ▷ No provably convergent algorithm.

Main Motivation: Generative Adversarial Networks (GANs)

- Objective of Wasserstein GANs:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} \left[D_w(X) \right] - \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} \left[D_w(X) \right].$$

  - ▷ θ: a **generator** neural net, generating fake data.
  - ▷ w: a **discriminator** neural net.
  - ▷ D_w: output of discriminator at w, *highly* non-convex/non-concave.

- Theoretical challenge:
  - ▷ A saddle point might NOT exist [Dasgupta and Maskin, 1986] (non-existence of pure strategy equilibrium).
  - ▷ No provably convergent algorithm.

- Practical challenge: An empirically working algorithm.
  - ▷ SGD does NOT work.
  - ▷ Adaptive methods (Adam, RMSProp, ...) are highly unstable, require tuning.

Goal of the Lecture: Generative Adversarial Networks (GANs)

- The Langevin-based approach:
  - ▷ Inspired by game theory, we lift the optimization problem to infinite-dimension:

pure strategy equilibria	$\rightarrow$	**mixed** strategy equilibria
finite dimension, non-convex/non-concave	$\rightarrow$	**infinite** dimension, **bi-affine**

Goal of the Lecture: Generative Adversarial Networks (GANs)

- The Langevin-based approach:
  - ▷ Inspired by game theory, we lift the optimization problem to infinite-dimension:

pure strategy equilibria	$\rightarrow$	**mixed** strategy equilibria
finite dimension, non-convex/non-concave	$\rightarrow$	**infinite** dimension, **bi-affine**

  - ▷ Provable, **infinite-dimensional** algorithms on space of probability measures.
  - ▷ A principled guideline for obtaining stable, but inefficient algorithms.
  - ▷ A simple heuristic for obtaining efficient, and *empirically* stable algorithms.

NOT our goal today: Game-Theoretical Learning

- In game-theoretical learning, we care about
  - ▷ Efficiently finding the NE.
  - ▷ Cumulative loss/regret for each player; i.e., you cannot suffer one player a lot just to find the NE.
  - ▷ Algorithmic challenge to adapt to the opponent (adversarial, collaborative, ...); best known results in [Kangarshahi* et al., 2018].

NOT our goal today: Game-Theoretical Learning

- In game-theoretical learning, we care about
  - ▷ Efficiently finding the NE.
  - ▷ Cumulative loss/regret for each player; i.e., you cannot suffer one player a lot just to find the NE.
  - ▷ Algorithmic challenge to adapt to the opponent (adversarial, collaborative, ...); best known results in [Kangarshahi* et al., 2018].

- In training GANs, we *only* care about finding the mixed NE.
  - ▷ The players (neural nets) are artificial; we do not care about their loss.
  - ▷ YOU design the algorithm; the environment is known.

Preliminary:
Bi-affine games with finite strategies

Bi-affine two-player games with finite strategies: The setting

- Bi-affine two-player games:

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} \langle y, a \rangle - \langle x, Ay \rangle \coloneqq \langle y^*, a \rangle - \langle x^*, Ay^* \rangle \qquad (\star)$$

  - ▷ Δ_d: d-simplex, **mixed** strategies over d pure strategies.
  - ▷ a, A: cost vector/matrix.
  - ▷ Goal: find a mixed Nash Equilibrium (NE) (x^*, y^*).

Bi-affine two-player games with finite strategies: The setting

- Bi-affine two-player games:

$$\min_{x\in\Delta_m}\max_{y\in\Delta_n} \langle y, a\rangle - \langle x, Ay\rangle \coloneqq \langle y^*, a\rangle - \langle x^*, Ay^*\rangle \qquad (\star)$$

  ▷ Δ_d: d-simplex, **mixed** strategies over d pure strategies.

  ▷ a, A: cost vector/matrix.

  ▷ Goal: find a mixed Nash Equilibrium (NE) (x^*, y^*).

- Existence of NE: von Neumann's minimax theorem [Neumann, 1928].

  ▷ $(\star)$ is a well-defined optimization problem.

Bi-affine two-player games with finite strategies: The setting

- Bi-affine two-player games:

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} \langle y, a \rangle - \langle x, Ay \rangle \coloneqq \langle y^*, a \rangle - \langle x^*, Ay^* \rangle \tag{$\star$}$$

  - ▷ Δ_d: d-simplex, **mixed** strategies over d pure strategies.
  - ▷ a, A: cost vector/matrix.
  - ▷ Goal: find a mixed Nash Equilibrium (NE) (x^*, y^*).

- Existence of NE: von Neumann's minimax theorem [Neumann, 1928].
  - ▷ ($\star$) is a well-defined optimization problem.

- Assumptions.
  - ▷ A is expensive to evaluate.
  - ▷ (Stochastic) gradients $a - A^\top x$ and $-Ay$ are cheap.

Bi-affine two-player games with finite strategies: algorithms

- Bi-affine two-player games:

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} \langle y, a \rangle - \langle x, Ay \rangle \coloneqq \langle y^*, a \rangle - \langle x^*, Ay^* \rangle$$

  - ▷ Δ_d: d-simplex, **mixed** strategies over d pure strategies
  - ▷ a, A: cost vector/matrix.

- Fundamental building blocks: Entropic Mirror Descent [Beck and Teboulle, 2003].
  - ▷ $\phi(z) = \sum_{i=1}^{d} z_i \log z_i,\ \phi^\star(z) = \log \sum_{i=1}^{d} e^{z_i}$.
  - ▷ MD iterates:

$$z' = \mathsf{MD}_\eta(z, b) \quad \equiv \quad z' = \nabla\phi^\star(\nabla\phi(z) - \eta b) \quad \equiv \quad z'_i = \frac{z_i e^{-\eta b_i}}{\sum_{i=1}^{d} z_i e^{-\eta b_i}}.$$

Bi-affine two-player games with finite strategies: algorithms

- Algorithm for finding NE: Entropic Mirror Descent and variants
  - ▷ Mirror descent [Beck and Teboulle, 2003]

$$\begin{cases} x_{t+1} = \mathsf{MD}_\eta\left(x_t, -A^\top y_t\right) \\ y_{t+1} = \mathsf{MD}_\eta\left(y_t, -a + Ax_t\right) \end{cases} \quad \Rightarrow \quad (\bar{x}_T, \bar{y}_T) \text{ is an } O\left(T^{-1/2}\right)\text{-NE.}$$

  - ▷ Mirror-Prox [Nemirovski, 2004]

$$\begin{cases} x_t = \mathsf{MD}_\eta\left(\tilde{x}_t, -A^\top \tilde{y}_t\right), & \tilde{x}_{t+1} = \mathsf{MD}_\eta\left(\tilde{x}_t, -A^\top y_t\right) \\ y_t = \mathsf{MD}_\eta\left(\tilde{y}_t, -a + A\tilde{x}_t\right), & \tilde{y}_{t+1} = \mathsf{MD}_\eta\left(\tilde{y}_t, -a + Ax_t\right) \end{cases}$$

$$\Rightarrow \quad (\bar{x}_T, \bar{y}_T) \text{ is an } O\left(T^{-1}\right)\text{-NE.}$$

Mixed NE of GANs:
Bi-affine games with **infinite** strategies

Wasserstein GANs, mixed Nash Equilibrium

- Objective of Wasseretein GANs, pure strategy equilibrium:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\mathsf{real}}} [D_w(X)] - \mathbb{E}_{X \sim P_{\mathsf{fake}}} [D_w(X)] \,.$$

Wasserstein GANs, mixed Nash Equilibrium

- Objective of Wasseretein GANs, pure strategy equilibrium:

$$\min_{\theta\in\Theta}\max_{w\in\mathcal{W}} \mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right] - \mathbb{E}_{X\sim P_{\text{fake}}}\left[D_w(X)\right].$$

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu\in\mathcal{M}(\Theta)}\max_{\mu\in\mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}\left[D_w(X)\right].$$

where $\mathcal{M}(\mathcal{Z}) := \{$all (regular) probability measures on $\mathcal{Z}\}$.

Wasserstein GANs, mixed Nash Equilibrium

- Objective of Wasseretein GANs, pure strategy equilibrium:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{X \sim P_{\text{fake}}} [D_w(X)] .$$

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{w \sim \mu} \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} [D_w(X)] .$$

where $\mathcal{M}(\mathcal{Z}) := \{\text{all (regular) probability measures on } \mathcal{Z}\}$.

- Existence of NE (ν^*, μ^*): Glicksberg's existence theorem [Glicksberg, 1952].
  - ▷ $\nu^* \neq \delta_{\theta^*}, \mu^* \neq \delta_{w^*}$
  - ▷ **Optimal strategies are indeed mixed !**

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{w \sim \mu} \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} [D_w(X)] . \qquad (\star)$$

- Define
  - ▷ (Riesz representation) $\langle \mu, h \rangle := \int h d\mu$ for a measure μ and function h.

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}\left[D_w(X)\right]. \qquad (\star)$$

- Define
  - ▷ (Riesz representation) $\langle \mu, h\rangle := \int h d\mu$ for a measure μ and function h.
  - ▷ Function $g(w) := \mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right]$

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu} \mathbb{E}_{X\sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}} [D_w(X)]. \qquad (\star)$$

- Define
  - ▷ (Riesz representation) $\langle \mu, h \rangle := \int h d\mu$ for a measure μ and function h.
  - ▷ Function $g(w) := \mathbb{E}_{X\sim P_{\text{real}}} [D_w(X)]$
  - ▷ Linear operator G and its adjoint $G^\dagger$:

$$G : \mathcal{M}(\Theta) \to \text{ a funciton on } \mathcal{W}, \quad G^\dagger : \mathcal{M}(\mathcal{W}) \to \text{ a funciton on } \Theta,$$
$$(G\nu)(w) := \mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}} [D_w(X)],$$
$$(G^\dagger\mu)(\theta) := \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}} [D_w(X)]$$

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]. \qquad (\star)$$

- Define
  - ▷ (Riesz representation) $\langle\mu, h\rangle := \int h d\mu$ for a measure μ and function h.
  - ▷ Function $g(w) := \mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)]$
  - ▷ Linear operator G and its adjoint $G^{\dagger}$:

$$G : \mathcal{M}(\Theta) \to \text{ a funciton on } \mathcal{W}, \quad G^{\dagger} : \mathcal{M}(\mathcal{W}) \to \text{ a funciton on } \Theta,$$
$$(G\nu)(w) := \mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)],$$
$$(G^{\dagger}\mu)(\theta) := \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

- Rewrite:

$$(\star)$$
$$\Updownarrow$$
$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \langle\mu, g\rangle - \langle\mu, G\nu\rangle$$

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}\left[D_w(X)\right]. \qquad (\star)$$

- Define
  - ▷ (Riesz representation) $\langle \mu, h\rangle := \int h d\mu$ for a measure μ and function h.
  - ▷ Function $g(w) := \mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right]$
  - ▷ Linear operator G and its adjoint $G^\dagger$:

$$G : \mathcal{M}(\Theta) \to \text{ a funciton on } \mathcal{W}, \quad G^\dagger : \mathcal{M}(\mathcal{W}) \to \text{ a funciton on } \Theta,$$
$$(G\nu)(w) := \mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}\left[D_w(X)\right],$$
$$(G^\dagger\mu)(\theta) := \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}\left[D_w(X)\right]$$

- Rewrite:

$$(\star)$$
$$\Updownarrow$$
$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \langle \mu, g\rangle - \langle \mu, G\nu\rangle$$

... a bi-affine game...

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Objective of Wasseretein GANs, **mixed** strategy equilibrium:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim P_{\text{real}}} \left[D_w(X) \right] - \mathbb{E}_{w \sim \mu} \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim P_{\text{fake}}^{\theta}} \left[D_w(X) \right]. \qquad (\star)$$

- Define
  - ▷ (Riesz representation) $\langle \mu, h \rangle := \int h d\mu$ for a measure μ and function h.
  - ▷ Function $g(w) := \mathbb{E}_{X \sim P_{\text{real}}} \left[D_w(X) \right]$
  - ▷ Linear operator G and its adjoint $G^\dagger$:

$$G : \mathcal{M}(\Theta) \to \text{ a funciton on } \mathcal{W}, \quad G^\dagger : \mathcal{M}(\mathcal{W}) \to \text{ a funciton on } \Theta,$$
$$(G\nu)(w) := \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim P_{\text{fake}}^{\theta}} \left[D_w(X) \right],$$
$$(G^\dagger \mu)(\theta) := \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim P_{\text{fake}}^{\theta}} \left[D_w(X) \right]$$

- Rewrite:

$$(\star)$$
$$\Updownarrow$$
$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle$$

... a bi-affine game... in infinite dimension!!

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Bi-affine two-player game, finite strategies:

$$\min_{x\in\Delta_m} \max_{y\in\Delta_n} \langle y, a\rangle - \langle x, Ay\rangle \,.$$

  - ▷ $\phi(z) = \sum_{i=1}^d z_i \log z_i,\ \phi^\star(z) = \log \sum_{i=1}^d e^{z_i}.$
  - ▷ Entropic MD iterates:

$$z' = \mathrm{MD}_\eta\,(z, b) \quad \equiv \quad z' = \nabla\phi^\star\left(\nabla\phi(z) - \eta b\right) \quad \equiv \quad z'_i = \frac{z_i e^{-\eta b_i}}{\sum_{i=1}^d z_i e^{-\eta b_i}}.$$

  - ▷ Entropic MD $\Rightarrow O(T^{-1/2})$-NE, MP $\Rightarrow O(T^{-1})$-NE.

- Bi-affine two-player game, **continuously infinite** strategies:

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \langle \mu, g\rangle - \langle \mu, G\nu\rangle \,.$$

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Bi-affine two-player game, finite strategies:

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} \langle y, a \rangle - \langle x, Ay \rangle .$$

  - ▷ $\phi(z) = \sum_{i=1}^d z_i \log z_i, \;\; \phi^\star(z) = \log \sum_{i=1}^d e^{z_i}.$
  - ▷ Entropic MD iterates:

$$z' = \mathrm{MD}_\eta(z, b) \quad \equiv \quad z' = \nabla\phi^\star(\nabla\phi(z) - \eta b) \quad \equiv \quad z'_i = \frac{z_i e^{-\eta b_i}}{\sum_{i=1}^d z_i e^{-\eta b_i}}.$$

  - ▷ Entropic MD $\Rightarrow O(T^{-1/2})$-NE, MP $\Rightarrow O(T^{-1})$-NE.

- Bi-affine two-player game, **continuously infinite** strategies:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle .$$

  - ▷ $\phi(z) = ??$, $\phi^\star(z) = ??$ (negative Shannon entropy, maybe?)

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Bi-affine two-player game, finite strategies:

$$\min_{x\in\Delta_m}\max_{y\in\Delta_n}\langle y,a\rangle-\langle x,Ay\rangle\,.$$

▷ $\phi(z)=\sum_{i=1}^d z_i\log z_i,\;\;\phi^\star(z)=\log\sum_{i=1}^d e^{z_i}.$

▷ Entropic MD iterates:

$$z'=\mathrm{MD}_\eta\left(z,b\right)\quad\equiv\quad z'=\nabla\phi^\star\left(\nabla\phi(z)-\eta b\right)\quad\equiv\quad z_i'=\frac{z_ie^{-\eta b_i}}{\sum_{i=1}^d z_ie^{-\eta b_i}}.$$

▷ Entropic MD $\Rightarrow O(T^{-1/2})$-NE, MP $\Rightarrow O(T^{-1})$-NE.

- Bi-affine two-player game, **continuously infinite** strategies:

$$\min_{\nu\in\mathcal{M}(\Theta)}\max_{\mu\in\mathcal{M}(\mathcal{W})}\langle\mu,g\rangle-\langle\mu,G\nu\rangle\,.$$

▷ $\phi(z)=$??, $\phi^\star(z)=$?? (negative Shannon entropy, maybe?)

▷ Entropic MD iterates in infinite dimension??

Wasserstein GANs, mixed Nash Equilibrium $\equiv$ bi-affine games

- Bi-affine two-player game, finite strategies:

$$\min_{x\in\Delta_m}\max_{y\in\Delta_n}\langle y,a\rangle-\langle x,Ay\rangle\,.$$

  ▷ $\phi(z)=\sum_{i=1}^d z_i\log z_i,\;\;\phi^\star(z)=\log\sum_{i=1}^d e^{z_i}.$

  ▷ Entropic MD iterates:

$$z'=\mathrm{MD}_\eta\left(z,b\right)\quad\equiv\quad z'=\nabla\phi^\star\left(\nabla\phi(z)-\eta b\right)\quad\equiv\quad z_i'=\frac{z_i e^{-\eta b_i}}{\sum_{i=1}^d z_i e^{-\eta b_i}}.$$

  ▷ Entropic MD $\Rightarrow O(T^{-1/2})$-NE, MP $\Rightarrow O(T^{-1})$-NE.

- Bi-affine two-player game, **continuously infinite** strategies:

$$\min_{\nu\in\mathcal{M}(\Theta)}\max_{\mu\in\mathcal{M}(\mathcal{W})}\langle\mu,g\rangle-\langle\mu,G\nu\rangle\,.$$

  ▷ $\phi(z)=??$, $\phi^\star(z)=??$ (negative Shannon entropy, maybe?)

  ▷ Entropic MD iterates in infinite dimension??

  ▷ Entropic MD in infinite dimension $\Rightarrow O(T^{-1/2})$-NE?? MP $\Rightarrow O(T^{-1})$-NE??

Entropic Mirror Descent Iterates in Infinite Dimension

- Negative Shannon entropy and its Fenchel dual: (dz :=Lebesgue)
  - ▷ $\Phi(\mu) = \int \mu \log \frac{d\mu}{dz}$.
  - ▷ $\Phi^\star(h) = \log \int e^h$.
  - ▷ $d\Phi$ and $d\Phi^\star$: Fréchet derivatives.[1]

Theorem (Infinite-Dimensional Mirror Descent, informal)

For a learning rate η, a probability measure μ, and an arbitrary function h, we can equivalently define

$$\mu_+ = \mathrm{MD}_\eta(\mu, h) \quad \equiv \quad \mu_+ = d\Phi^\star\left(d\Phi(\mu) - \eta h\right) \equiv \quad d\mu_+ = \frac{e^{-\eta h} d\mu}{\int e^{-\eta h} d\mu}.$$

Moreover, most the essential ingredients in the analysis of finite-dimensional prox methods can be generalized to infinite dimension.

[1] Under mild regularity conditions on the measure/function.

Entropic Mirror Descent Iterates in Infinite Dimension

- Negative Shannon entropy and its Fenchel dual: (dz :=Lebesgue)
  - ▷ $\Phi(\mu) = \int \mu \log \frac{d\mu}{dz}$.
  - ▷ $\Phi^\star(h) = \log \int e^h$.
  - ▷ $d\Phi$ and $d\Phi^\star$: Fréchet derivatives.[1]

Theorem (Infinite-Dimensional Mirror Descent, informal)

For a learning rate η, a probability measure μ, and an arbitrary function h, we can equivalently define

$$\mu_+ = \mathrm{MD}_\eta(\mu, h) \quad \equiv \quad \mu_+ = d\Phi^\star\left(d\Phi(\mu) - \eta h\right) \equiv \quad d\mu_+ = \frac{e^{-\eta h} d\mu}{\int e^{-\eta h} d\mu}.$$

Moreover, most the essential ingredients in the analysis of finite-dimensional prox methods can be generalized to infinite dimension.

In particular, the three-point identity holds: For all μ, μ', μ'',

$$\left\langle \mu'' - \mu, d\Phi(\mu') - d\Phi(\mu) \right\rangle = D_\Phi(\mu, \mu') + D_\Phi(\mu'', \mu) - D_\Phi(\mu'', \mu').$$

where D_Φ is the Bregman divergence associated with Φ.

[1]Under mild regularity conditions on the measure/function.

Entropic Mirror Descent/Mirror-Prox in Infinite Dimension: Rates

- Algorithms:

Algorithm 1: Infinite-Dimensional Entropic MD

Input: Initial distributions μ_1, ν_1, learning rate η

for $t = 1, 2, \ldots, T-1$ **do**

$\quad \nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

Algorithm 2: Infinite-Dimensional Entropic MP

Input: Initial distributions $\tilde{\mu}_1, \tilde{\nu}_1$, learning rate η

for $t = 1, 2, \ldots, T$ **do**

$\quad \nu_t = \mathrm{MD}_\eta\left(\tilde{\nu}_t, -G^\dagger \tilde{\mu}_t\right), \quad \mu_t = \mathrm{MD}_\eta\left(\tilde{\mu}_t, -g + G\tilde{\nu}_t\right);$

$\quad \tilde{\nu}_{t+1} = \mathrm{MD}_\eta\left(\tilde{\nu}_t, -G^\dagger \mu_t\right), \quad \tilde{\mu}_{t+1} = \mathrm{MD}_\eta\left(\tilde{\mu}_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

Entropic Mirror Descent/Mirror-Prox in Infinite Dimension: Rates

- Algorithms:

Algorithm 1: Infinite-Dimensional Entropic MD

Input: Initial distributions μ_1, ν_1, learning rate η

for $t = 1, 2, \ldots, T-1$ **do**

$\quad \nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

Algorithm 2: Infinite-Dimensional Entropic MP

Input: Initial distributions $\tilde{\mu}_1, \tilde{\nu}_1$, learning rate η

for $t = 1, 2, \ldots, T$ **do**

$\quad \nu_t = \mathrm{MD}_\eta\left(\tilde{\nu}_t, -G^\dagger \tilde{\mu}_t\right), \quad \mu_t = \mathrm{MD}_\eta\left(\tilde{\mu}_t, -g + G\tilde{\nu}_t\right);$

$\quad \tilde{\nu}_{t+1} = \mathrm{MD}_\eta\left(\tilde{\nu}_t, -G^\dagger \mu_t\right), \quad \tilde{\mu}_{t+1} = \mathrm{MD}_\eta\left(\tilde{\mu}_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

Theorem (Convergence Rates)

Let $\Phi(\mu) = \int d\mu \log \frac{d\mu}{dz}$. *Then*

1. *Entropic MD* $\Rightarrow O(T^{-1/2})$*-NE, MP* $\Rightarrow O(T^{-1})$*-NE.*
2. *If only stochastic derivatives* $(-\hat{G}^\dagger \mu$ *and* $\hat{g} - \hat{G}\nu)$ *are available, then both MD and MP* $\Rightarrow O(T^{-1/2})$*-NE in expectation.*

Summary so far

- Mixed NE of GANs $\equiv$ bi-affine two-player game with **continuously infinite** strategies:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle$$

- Can find NE of GANs via entropic MD and MP with rigorous rates, provided...
  - ▷ Have access to derivatives $-G^\dagger \mu$ and $g - G\nu$.
  - ▷ Can update probability measures $\mu_+ = \text{MD}_\eta\,(\mu, \text{derivative})$.

Summary so far

- Mixed NE of GANs $\equiv$ bi-affine two-player game with **continuously infinite** strategies:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle$$

- Can find NE of GANs via entropic MD and MP with rigorous rates, provided...
  - ▷ Have access to derivatives $-G^{\dagger}\mu$ and $g - G\nu$.
  - ▷ Can update probability measures $\mu_{+} = \mathrm{MD}_{\eta}\left(\mu, \text{derivative}\right)$.
  - ▷ ...but we can do neither.

Summary so far

- Mixed NE of GANs $\equiv$ bi-affine two-player game with **continuously infinite** strategies:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle$$

- Can find NE of GANs via entropic MD and MP with rigorous rates, provided...
  - ▷ Have access to derivatives $-G^\dagger \mu$ and $g - G\nu$.
  - ▷ Can update probability measures $\mu_+ = \mathrm{MD}_\eta\left(\mu, \text{derivative}\right)$.
  - ▷ ...but we can do neither.

- Solution to both: **Sampling**.

Summary so far

- Mixed NE of GANs $\equiv$ bi-affine two-player game with **continuously infinite** strategies:

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle$$

- Can find NE of GANs via entropic MD and MP with rigorous rates, provided...
  - ▷ Have access to derivatives $-G^\dagger \mu$ and $g - G\nu$.
  - ▷ Can update probability measures $\mu_+ = \mathrm{MD}_\eta\,(\mu, \text{derivative})$.
  - ▷ ...but we can do neither.

- Solution to both: **Sampling**.
  - ▷ Empirical estimate for stochastic derivatives.
  - ▷ Draw samples from the updated probability measure via **Langevin dynamics**.

From Theory to Practice: Sampling via Langevin Dynamics

Reduction from Entropic MD to Sampling

- Goal of the section:
  - ▷ A principled reduction from entropic MD to sampling.[2]

Algorithm 1: Infinite-Dimensional Entropic MD

Input: Initial distributions μ_1, ν_1, learning rate η

for $t = 1, 2, \dots, T-1$ **do**

$\quad \nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

  - ▷ A heuristic for reducing computational/memory burdens.

- We set $\eta = 1$ since it plays no role in the following derivation.

[2] Entropic Mirror-Prox can be similarly reduced.

Reduction Step 1: Reformulating MD Iterates

Lemma (Iterates of Entropic MD)

Entropic MD is equivalent to:

$$d\mu_T = \frac{\exp\left\{(T-1)g - G\sum_{s=1}^{T-1}\nu_s\right\}dw}{\int \exp\left\{(T-1)g - G\sum_{s=1}^{T-1}\nu_s\right\}dw}.$$

and

$$d\nu_T = \frac{\exp\left\{G^\dagger\sum_{s=1}^{T-1}\mu_s\right\}d\theta}{\int \exp\left\{G^\dagger\sum_{s=1}^{T-1}\mu_s\right\}d\theta}.$$

Proof.

A well-known result in finite dimension; the proof holds verbatim given our infinite-dimensional framework for MD. □

Reduction Step 1: Reformulating MD Iterates

Lemma (Iterates of Entropic MD)

Entropic MD is equivalent to:

$$d\mu_T = \frac{\exp\left\{(T-1)g - G\sum_{s=1}^{T-1}\nu_s\right\}dw}{\int \exp\left\{(T-1)g - G\sum_{s=1}^{T-1}\nu_s\right\}dw}.$$

and

$$d\nu_T = \frac{\exp\left\{G^\dagger \sum_{s=1}^{T-1}\mu_s\right\}d\theta}{\int \exp\left\{G^\dagger \sum_{s=1}^{T-1}\mu_s\right\}d\theta}.$$

Proof.

A well-known result in finite dimension; the proof holds verbatim given our infinite-dimensional framework for MD. □

- Implication:

Want to sample from $\mu_T/\nu_T \equiv$ Need (stochastic) derivatives of history.

Reduction Step 2: Stochastic Derivatives based on Samples

- Empirical approximation for stochastic derivatives
  - ▷ Suppose we can acquire samples $\{\theta_i\}_{i=1}^n \sim \nu_1$,

$$(G\nu_1)(w) = \mathbb{E}_{\theta\sim\nu_1}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta_i}}[f_w(X)] \coloneqq \hat{G}\nu_1(w)$$

Reduction Step 2: Stochastic Derivatives based on Samples

- Empirical approximation for stochastic derivatives
  - ▷ Suppose we can acquire samples $\{\theta_i\}_{i=1}^n \sim \nu_1$,

$$(G\nu_1)(w) = \mathbb{E}_{\theta\sim\nu_1}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P^{\theta_i}_{\text{fake}}}[f_w(X)] \coloneqq \hat{G}\nu_1(w)$$

  - ▷ Similarly, for $\{w_i\}_{i=1}^n \sim \mu_1$,

$$(G^\dagger\mu_1)(\theta) = \mathbb{E}_{w\sim\mu_1}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[f_{w_i}(X)] \coloneqq \hat{G}^\dagger\mu_1(\theta)$$

Reduction Step 2: Stochastic Derivatives based on Samples

- Empirical approximation for stochastic derivatives
  - ▷ Suppose we can acquire samples $\{\theta_i\}_{i=1}^n \sim \nu_1$,

$$(G\nu_1)(w) = \mathbb{E}_{\theta\sim\nu_1}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X\sim P^{\theta_i}_{\text{fake}}}[f_w(X)] \coloneqq \hat{G}\nu_1(w)$$

  - ▷ Similarly, for $\{w_i\}_{i=1}^n \sim \mu_1$,

$$(G^\dagger\mu_1)(\theta) = \mathbb{E}_{w\sim\mu_1}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[f_{w_i}(X)] \coloneqq \hat{G}^\dagger\mu_1(\theta)$$

- Key question:

Given samples from (μ_1, ν_1), how do we draw samples from (μ_2, ν_2)?

Reduction Step 2: Stochastic Derivatives based on Samples

- Empirical approximation for stochastic derivatives
  - ▷ Suppose we can acquire samples $\{\theta_i\}_{i=1}^n \sim \nu_1$,

$$(G\nu_1)(w) = \mathbb{E}_{\theta\sim\nu_1}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta_i}}[f_w(X)] \coloneqq \hat{G}\nu_1(w)$$

  - ▷ Similarly, for $\{w_i\}_{i=1}^n \sim \mu_1$,

$$(G^\dagger\mu_1)(\theta) = \mathbb{E}_{w\sim\mu_1}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}[f_w(X)] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P_{\text{fake}}^{\theta}}[f_{w_i}(X)] \coloneqq \hat{G}^\dagger\mu_1(\theta)$$

- Key question:

Given samples from (μ_1, ν_1), how do we draw samples from (μ_2, ν_2)?

  - ▷ If we can do this, then applying recursively, we can get samples from (μ_T, ν_T).

Reduction Step 3: Stochastic Gradient Langevin Dynamics (SGLD)

- Given any distribution $d\mu = \frac{e^{-h}}{\int e^{-h}}$, the SGLD iterates

$$z_{k+1} = z_k - \gamma \hat{\nabla} h(z_k) + \sqrt{2\gamma} \epsilon \mathcal{N}(0, I)$$

  - ▷ γ step-size, $\hat{\nabla}$ stochastic gradients, ϵ thermal noise.

Reduction Step 3: Stochastic Gradient Langevin Dynamics (SGLD)

- Given any distribution $d\mu = \frac{e^{-h}}{\int e^{-h}}$, the SGLD iterates

$$z_{k+1} = z_k - \gamma \hat{\nabla} h(z_k) + \sqrt{2\gamma}\epsilon \mathcal{N}(0, I)$$

  - ▷ γ step-size, $\hat{\nabla}$ stochastic gradients, ϵ thermal noise.
  - ▷ z_k converges to samples from μ [Welling and Teh, 2011].

Reduction Step 3: Stochastic Gradient Langevin Dynamics (SGLD)

- Given any distribution $d\mu = \frac{e^{-h}}{\int e^{-h}}$, the SGLD iterates

$$z_{k+1} = z_k - \gamma \hat{\nabla} h(z_k) + \sqrt{2\gamma}\epsilon \mathcal{N}(0, I)$$

  - ▷ γ step-size, $\hat{\nabla}$ stochastic gradients, ϵ thermal noise.
  - ▷ z_k converges to samples from μ [Welling and Teh, 2011].
  - ▷ Used in deep learning, empirically successful [Chaudhari et al., 2017, Dziugaite and Roy, 2018, Chaudhari et al., 2018].

Reduction Step 3: Stochastic Gradient Langevin Dynamics (SGLD)

- Given any distribution $d\mu = \frac{e^{-h}}{\int e^{-h}}$, the SGLD iterates

$$z_{k+1} = z_k - \gamma \hat{\nabla} h(z_k) + \sqrt{2\gamma}\epsilon \mathcal{N}(0, I)$$

  - ▷ γ step-size, $\hat{\nabla}$ stochastic gradients, ϵ thermal noise.
  - ▷ z_k converges to samples from μ [Welling and Teh, 2011].
  - ▷ Used in deep learning, empirically successful [Chaudhari et al., 2017, Dziugaite and Roy, 2018, Chaudhari et al., 2018].

- Let $h \leftarrow -\hat{G}^\dagger \mu_1$, then $\nu_2 \propto e^{-h}$.

Reduction Step 3: Stochastic Gradient Langevin Dynamics (SGLD)

- Given any distribution $d\mu = \frac{e^{-h}}{\int e^{-h}}$, the SGLD iterates

$$z_{k+1} = z_k - \gamma \hat{\nabla} h(z_k) + \sqrt{2\gamma} \epsilon \mathcal{N}(0, I)$$

  - ▷ γ step-size, $\hat{\nabla}$ stochastic gradients, ϵ thermal noise.
  - ▷ z_k converges to samples from μ [Welling and Teh, 2011].
  - ▷ Used in deep learning, empirically successful [Chaudhari et al., 2017, Dziugaite and Roy, 2018, Chaudhari et al., 2018].

- Let $h \leftarrow -\hat{G}^\dagger \mu_1$, then $\nu_2 \propto e^{-h}$.
  - ▷ SGLD requires stochastic gradients of h. We again use empirical approximation:

$$\nabla_\theta h = \nabla_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{X \sim P^\theta_{\text{fake}}} [f_{w_i}(X)] \right) \simeq \nabla_\theta \left(\frac{1}{nn'} \sum_{i,j=1}^{n} f_{w_i}(X_j) \right) \coloneqq \hat{\nabla}_\theta h,$$

  - ▷ $X_j \sim P^\theta_{\text{fake}}$ for $j = 1, 2, ..., n'$.

Implementable Entropic Mirror Descent

- Combining Steps 1-3, we get an implementable algorithm [Hsieh et al., 2019].

- Caveat: The algorithm is **not** practical since
  - ▷ The algorithm is complicated.
  - ▷ The algorithm takes $O(T)$ memory and $O(T^2)$ per-iteration complexity since we need to memorize all history.

Implementable Entropic Mirror Descent

- Combining Steps 1-3, we get an implementable algorithm [Hsieh et al., 2019].

- Caveat: The algorithm is **not** practical since
  - ▷ The algorithm is complicated.
  - ▷ The algorithm takes $O(T)$ memory and $O(T^2)$ per-iteration complexity since we need to memorize all history.

- Key question:

  How to reduce the complexity while maintaining the sampling perspective?

⋆ Efficient Entropic Mirror Descent

- Our proposal: Summarizing the samples by their mean.

$$\{\theta_i\}_{i=1}^n \quad \rightsquigarrow \quad \bar{\theta} \coloneqq \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \alpha_i \theta_i \quad (\simeq \mathbb{E}_{\theta\sim\nu}[\theta])\,,$$

$$\{w_i\}_{i=1}^n \quad \rightsquigarrow \quad \bar{w} \coloneqq \frac{1}{\sum_{i=1}^n \alpha_i'} \sum_{i=1}^n \alpha_i' w_i \quad (\simeq \mathbb{E}_{w\sim\mu}[w])\,.$$

⋆ Efficient Entropic Mirror Descent

- Our proposal: Summarizing the samples by their mean.

$$\{\theta_i\}_{i=1}^n \quad \rightsquigarrow \quad \bar{\theta} \coloneqq \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \alpha_i \theta_i \quad \left(\simeq \mathbb{E}_{\theta\sim\nu}[\theta]\right),$$

$$\{w_i\}_{i=1}^n \quad \rightsquigarrow \quad \bar{w} \coloneqq \frac{1}{\sum_{i=1}^n \alpha_i'} \sum_{i=1}^n \alpha_i' w_i \quad \left(\simeq \mathbb{E}_{w\sim\mu}[w]\right).$$

- Resulting algorithm: Mirror-GAN
  - ▷ Linear time.
  - ▷ Constant memory.
  - ▷ Complexity $\simeq$ SGD.

⋆ Efficient Entropic Mirror Descent

- Our proposal: Summarizing the samples by their mean.

$$\{\theta_i\}_{i=1}^n \quad \rightsquigarrow \quad \bar{\theta} \coloneqq \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \alpha_i \theta_i \quad \left(\simeq \mathbb{E}_{\theta\sim\nu}[\theta]\right),$$

$$\{w_i\}_{i=1}^n \quad \rightsquigarrow \quad \bar{w} \coloneqq \frac{1}{\sum_{i=1}^n \alpha'_i} \sum_{i=1}^n \alpha'_i w_i \quad \left(\simeq \mathbb{E}_{w\sim\mu}[w]\right).$$

- Resulting algorithm: Mirror-GAN
  - ▷ Linear time.
  - ▷ Constant memory.
  - ▷ Complexity $\simeq$ SGD.

Remark: In principle, we can use any first-order algorithm to compute the mean, not necessarily SGLD.

$\star$ Efficient Entropic Mirror Descent: Algorithm [Hsieh et al., 2018b]

Algorithm 3: MIRROR-GAN: APPROXIMATE MIRROR DECENT FOR GANS

Input: $\bar{\boldsymbol{w}}_1, \bar{\boldsymbol{\theta}}_1 \leftarrow$ random initialization, $\{\gamma_t\}_{t=1}^T, \{\epsilon_t\}_{t=1}^T, \{K_t\}_{t=1}^{T-1}, \beta$ (see Appendix D for meaning of the hyperparameters).

for $t = 1, 2, \ldots, T-1$ **do**

- $\bar{\boldsymbol{w}}_t, \boldsymbol{w}_t^{(1)} \leftarrow \boldsymbol{w}_t$;
- $\bar{\boldsymbol{\theta}}_t, \boldsymbol{\theta}_t^{(1)} \leftarrow \boldsymbol{\theta}_t$;
- **for** $k = 1, 2, \ldots, K_t$ **do**
  - Generate $A = \{X_1, \ldots, X_n\} \sim \mathbb{P}_{\boldsymbol{\theta}_t^{(k)}}$;
  - $\boldsymbol{\theta}_t^{(k+1)} = \boldsymbol{\theta}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{\theta}} \sum_{X_i \in A} f_{\boldsymbol{w}_t}(X_i) + \sqrt{2\gamma_t}\epsilon_t \mathcal{N}(0, I)$;
  - Generate $B = \{X_1^{\text{real}}, \ldots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$;
  - Generate $B' = \{X_1', \ldots, X_n'\} \sim \mathbb{P}_{\boldsymbol{\theta}_t}$;

    $$\boldsymbol{w}_t^{(k+1)} = \boldsymbol{w}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{w}} \sum_{X_i^{\text{real}} \in B} f_{\boldsymbol{w}_t^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{n} \nabla_{\boldsymbol{w}} \sum_{X_i' \in B'} f_{\boldsymbol{w}_t^{(k)}}(X_i') + \sqrt{2\gamma_t}\epsilon_t \mathcal{N}(0, I);$$

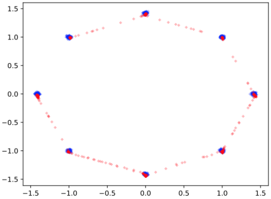
  - $\bar{\boldsymbol{w}}_t \leftarrow (1-\beta)\bar{\boldsymbol{w}}_t + \beta \boldsymbol{w}_t^{(k+1)}$;
  - $\bar{\boldsymbol{\theta}}_t \leftarrow (1-\beta)\bar{\boldsymbol{\theta}}_t + \beta \boldsymbol{\theta}_t^{(k+1)}$;
- $\boldsymbol{w}_{t+1} \leftarrow (1-\beta)\boldsymbol{w}_t + \beta \bar{\boldsymbol{w}}_t$;
- $\boldsymbol{\theta}_{t+1} \leftarrow (1-\beta)\boldsymbol{\theta}_t + \beta \bar{\boldsymbol{\theta}}_t$;

return $\boldsymbol{w}_T, \boldsymbol{\theta}_T$.

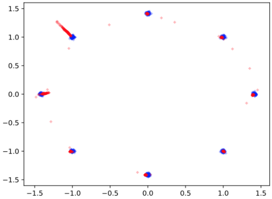
Experiments

Synthetic Data: 8 Gaussian mixtures

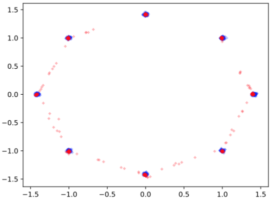
- Observations:
  - ▷ Mirror-/Mirror-Prox-GAN capture mode/variance more accurately.
  - ▷ Mirror-/Mirror-Prox-GAN **never** overfit (mode collapse), in contrast to Adam.



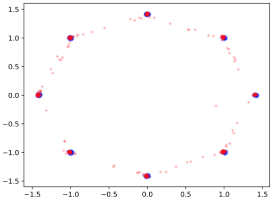
(a) SGD



(b) Adam

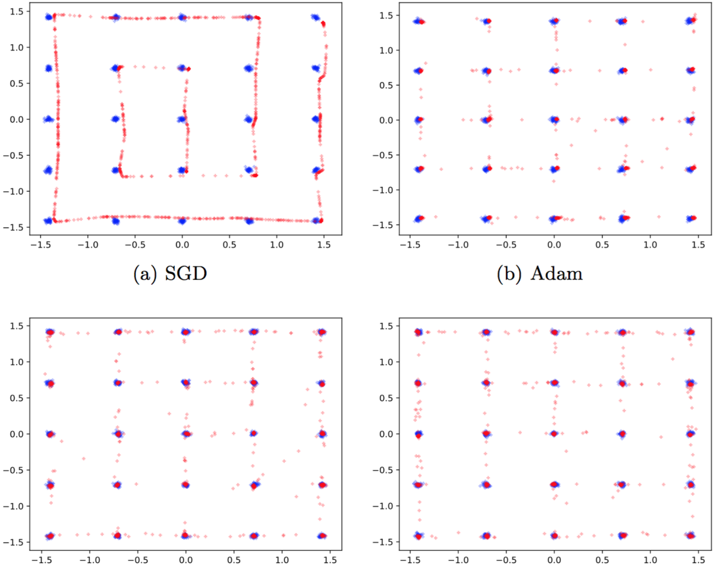


(c) Mirror-GAN



(d) Mirror-Prox-GAN

Synthetic Data: 25 Gaussian mixtures

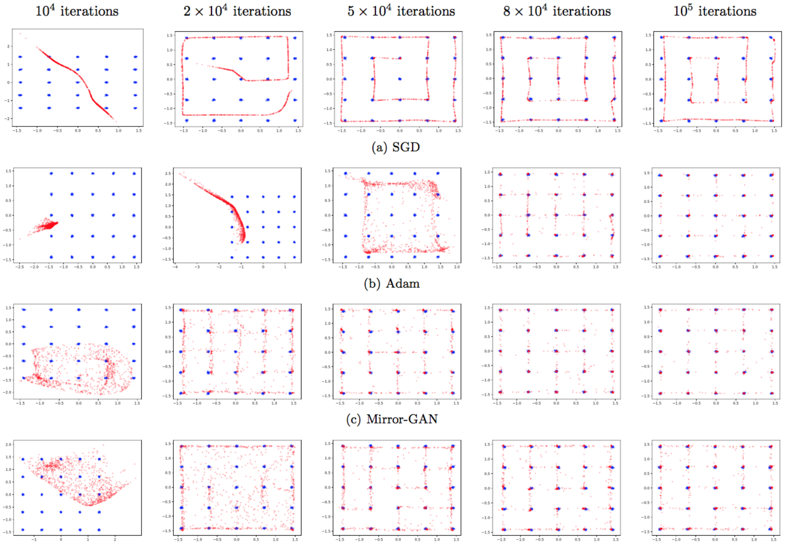


(a) SGD (b) Adam

(c) Mirror-GAN (d) Mirror-Prox-GAN

Synthetic Data: 25 Gaussian mixtures (cont.)

- Mirror-/Mirror-Prox-GAN are faster.



(a) SGD

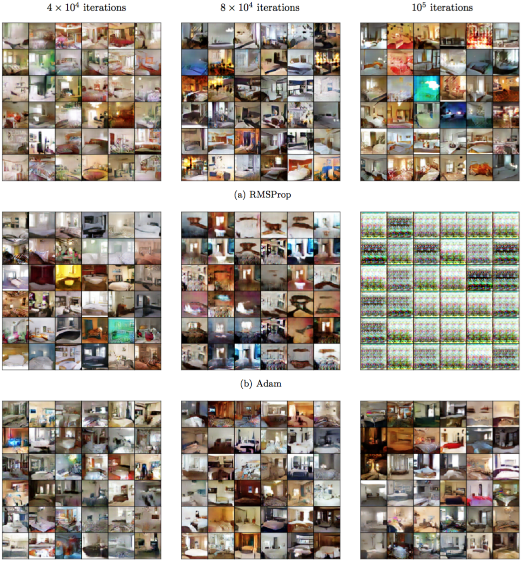
(b) Adam

(c) Mirror-GAN

(d) Mirror-Prox-GAN

Real Data

- Better image quality.



(a) RMSProp

(b) Adam

(c) Mirror-GAN, **Algorithm 3**

Summary and potential follow-ups

- Re-think the framework for GANs.
  - ▷ From **pure strategy** equilibria to **mixed** Nash Equilibira.
  - ▷ Mixed Nash Equilibira of GANs are but bi-affine two-player games!!!
  - ▷ Infinite-dimensional mirror descent/mirror-prox provably find the mixed NE.

- Algorithmic approach.
  - ▷ Sampling for stochastic derivatives and stochastic gradients.
  - ▷ Summarizing samples by mean; efficient algorithms.

Summary and potential follow-ups

- Re-think the framework for GANs.
  - ▷ From **pure strategy** equilibria to **mixed** Nash Equilibira.
  - ▷ Mixed Nash Equilibira of GANs are but bi-affine two-player games!!!
  - ▷ Infinite-dimensional mirror descent/mirror-prox provably find the mixed NE.

- Algorithmic approach.
  - ▷ Sampling for stochastic derivatives and stochastic gradients.
  - ▷ Summarizing samples by mean; efficient algorithms.

- Future work.
  - ▷ We have only considered WGANs, but our framework applies to any GAN.
  - ▷ More generally, can apply our framework to any min-max objective (e.g., adversarial training, robust reinforcement learning).

Langevin dynamics for non-convex optimization

Need for noise in non-convex optimization

- Standard unconstrained optimization problem:

$$\max_{\mathbf{x}\in\mathbb{R}^d} V(\mathbf{x}) \tag{2}$$

- Global optimization of non-convex objective:
  - ▷ Escape saddle-points
  - ▷ Escape local maxima

- Proposed solution: add noise to the iterates $\rightarrow$ Langevin Dynamics !

- Studies of LD for non-convex optimization [Xu et al., 2018, Gao et al., 2018]

A sampling perspective of optimization

- Standard unconstrained optimization problem:

$$\max_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \tag{3}$$

- Similarly as with GANs, (3) can be recasted as

$$\max_{\mu \in \mathcal{M}(\mathbb{R}^d)} \mathbb{E}_{\mathbf{x} \sim \mu}[f(\mathbf{x})] \tag{4}$$

  ▷ The exact solution $\mu^\star = \delta_{x^\star}$ is a Dirac delta function δ on the solution $x^\star$ to (3).

- Add entropy regularizer:

$$\max_{\mu \in \mathcal{M}(\mathbb{R}^d)} \mathbb{E}_{\mathbf{x} \sim \mu}[f(\mathbf{x})] + \alpha \mathcal{H}(\mu) \tag{5}$$

where $\alpha \in \mathbb{R}^+$ and $\mathcal{H}(\mu) \equiv \mathbb{E}_{\mathbf{x} \sim \mu}[-\log \mu(\mathbf{x})]$.

  ▷ The exact solution becomes $\mu_\alpha^*(\mathbf{x}) \propto e^{\frac{1}{\alpha} f(\mathbf{x})} \mathrm{d}\mathbf{x} \xrightarrow[\alpha \to 0]{} \delta_{x^*}$

Preconditioned SGLD

- Sample from μ_α^* using SGLD.

- Speed up sampling using preconditioning matrix $G(\mathbf{x})$ [Patterson and Teh, 2013]:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \beta_k \left(\frac{1}{\alpha} G(\mathbf{x}^k) \nabla f(\mathbf{x}^k) + \Gamma(\mathbf{x}^k) \right) + \sqrt{2\beta_k} G(\mathbf{x}^k)^{\frac{1}{2}} \boldsymbol{\xi}^k, \tag{6}$$

where $\Gamma(\mathbf{x})_i = \sum_{j=1}^d \frac{\partial G(\mathbf{x})_{ij}}{\partial \mathbf{x}_j}$.

- Good choice of preconditioning matrix: inverse Fisher Information Matrix.

- Diagonal approximation of inverse FIM: RMSProp preconditioner

$$G(\mathbf{x}^{k+1}) = \text{diag}\left(\mathbf{1} \oslash \left(\lambda \mathbf{1} + \sqrt{\bar{g}(\mathbf{x}^{k+1})} \right) \right)$$
$$\bar{g}(\mathbf{x}^{k+1}) = \gamma \bar{g}(\mathbf{x}^k) + \frac{1-\gamma}{\alpha^2} \nabla f(\mathbf{x}^k) \odot \nabla f(\mathbf{x}^k)$$

where $\gamma \in [0,1]$, and operators $\oslash, \odot$ represent the element-wise matrix product and division, respectively.

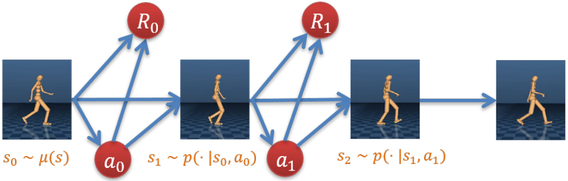
- When γ is small, the $\Gamma(\mathbf{x}^k)$ term in (6) can be safely ignored.

SGLD algorithm for training Deep Neural Networks [Li et al., 2016]

Algorithm 1 Preconditioned SGLD with RMSprop

Inputs: $\{\epsilon_t\}_{t=1:T}, \lambda, \alpha$
Outputs: $\{\boldsymbol{\theta}_t\}_{t=1:T}$
Initialize: $\mathbf{V}_0 \leftarrow \mathbf{0}$, random $\boldsymbol{\theta}_1$
for $t \leftarrow 1 : T$ **do**
  Sample a minibatch of size n, $\mathcal{D}_n^t = \{\boldsymbol{d}_{t_1}, \ldots, \boldsymbol{d}_{t_n}\}$
  Estimate gradient $\bar{g}(\boldsymbol{\theta}_t; \mathbf{X}^t) = \frac{1}{n}\sum_{i=1}^{n} \nabla \log p(\boldsymbol{d}_{t_i}|\boldsymbol{\theta}_t)$
  $V(\boldsymbol{\theta}_t) \leftarrow \alpha V(\boldsymbol{\theta}_{t-1}) + (1-\alpha)\bar{g}(\boldsymbol{\theta}_t; \mathcal{D}^t) \odot \bar{g}(\boldsymbol{\theta}_t; \mathcal{D}^t)$
  $G(\boldsymbol{\theta}_t) \leftarrow \mathrm{diag}\left(\mathbf{1} \oslash \left(\lambda \mathbf{1} + \sqrt{V(\boldsymbol{\theta}_t)}\right)\right)$
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \frac{\epsilon_t}{2}\left[G(\boldsymbol{\theta}_t)\left(\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}_t) + N\bar{g}(\boldsymbol{\theta}_t; \mathcal{D}^t)\right) + \Gamma(\boldsymbol{\theta}_t)\right] + \mathcal{N}(0, \epsilon_t G(\boldsymbol{\theta}_t))$
end for

Reinforcement Learning: Setup



- Parametrized deterministic policy: $\pi_\theta : \mathcal{S} \to \mathcal{A}$.
- Goal:

$$\max_\theta J(\theta) := \mathbb{E}_{s_0, r_1, s_1, \ldots} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, a_i = \pi_\theta(s_i), i = 1, 2, \ldots \right]$$

SGLD for training a Reinforcement Learning agent

- Policy Gradient Theorem:

$$\nabla J(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta}} \left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s,a) \,|_{a=\pi_\theta(s)} \right],$$

▷ $Q^{\pi_\theta}(s,a) = \mathbb{E}_{\pi_\theta} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| S_t = s, A_t = a \right]$,

▷ d^{π_θ} is the stationary state distribution under π_θ.

▷ Can (almost) be treated as a standard first order optimization problem.

▷ Non-convex objective: need for exploration.

- Other approaches [Lillicrap et al., 2015, Plappert et al., 2017]

▷ Add noise during the episode collection phase

- SGLD: directly adds noise when updating the parameters

▷ Annealing temperature: $\alpha_k \propto k^{-r}$, $(r = 0.25)$.

Empirical comparison between various exploration methods

- A comparison of exploration strategies with TD3 algorithm [Fujimoto et al., 2018]

  ▷ Action: adds noise to actions during the episode collection phase

  ▷ Parameter: adds noise the policy parametrs during the episode collection phase

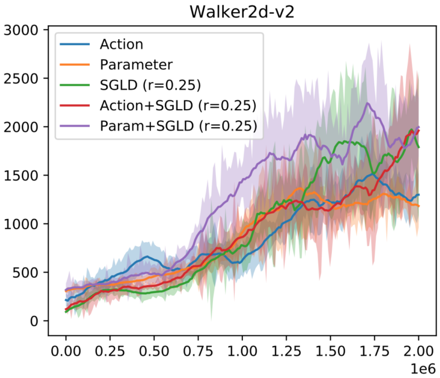
  ▷ SGLD: adds noise during parameter update

  ▷ SGLD+Action: adds noise both to action during episode collection and during parameter update phase

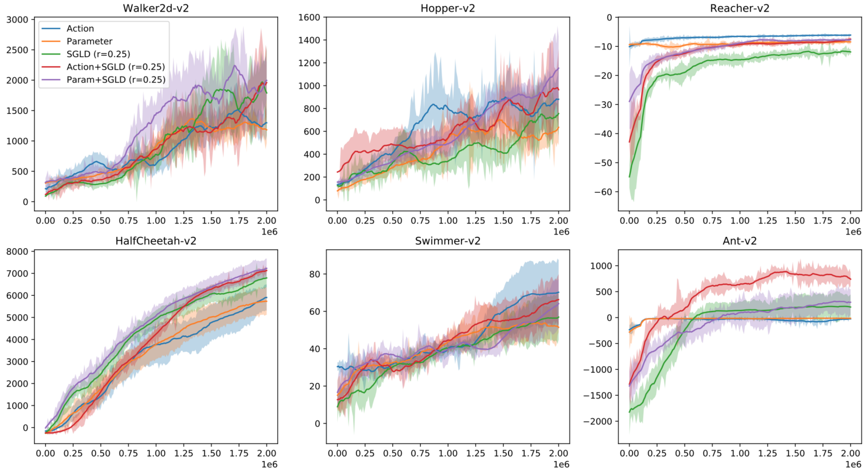
  ▷ SGLD+Parameter: adds noise both to parameters during episode collection and during parameter update phase

Expertiment on Walker2d Mujoco environment

- Expected reward averaged over 10 episodes
- Shaded area: half standard deviation
- For each algorithm, we optimize the hyperparameters, i.e.:
  - ▷ initial step-size
  - ▷ noise level
- Improved efficiency of Langevin methods

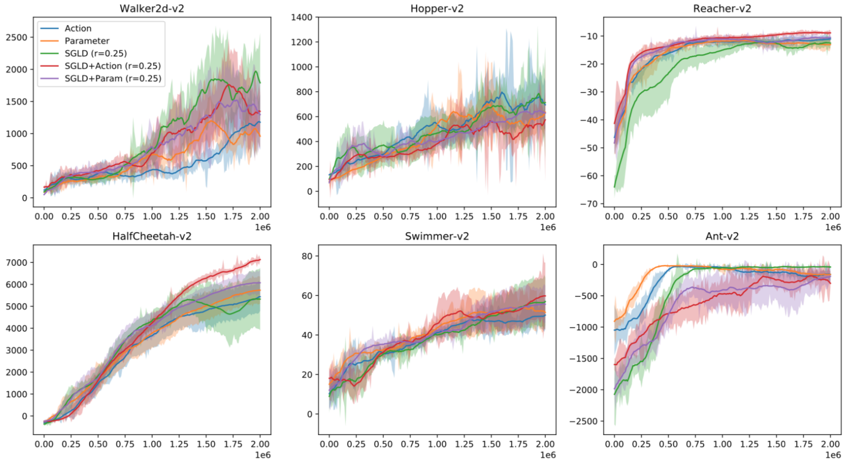


Experiments on several Mujoco environments



Generalization across various environments

- Generalization: Fixed set of hyper-parameters across all environments



That's all folks!!!

- https://lions.epfl.ch/publications
- https://arxiv.org/abs/1802.10174

[0] Beck, A. and Teboulle, M. (2003).
Mirror descent and nonlinear projected subgradient methods for convex optimization.
Operations Research Letters, 31(3):167–175.

[0] Bernton, E. (2018).
Langevin monte carlo and jko splitting.
arXiv preprint arXiv:1802.08671.

[0] Brosse, N., Durmus, A., Moulines, É., and Pereyra, M. (2017).
Sampling from a log-concave distribution with compact support with proximal langevin monte carlo.
arXiv preprint arXiv:1705.08964.

[0] Bubeck, S., Eldan, R., and Lehec, J. (2015).
Sampling from a log-concave distribution with projected langevin monte carlo.
arXiv preprint arXiv:1507.02564.

[0] Chambolle, A. and Pock, T. (2011).
A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of mathematical imaging and vision, 40(1):120–145.

[0] Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. (2017).
Entropy-sgd: Biasing gradient descent into wide valleys.
In *International Conference on Learning Representations*.

[0] Chaudhari, P., Oberman, A., Osher, S., Soatto, S., and Carlier, G. (2018).
Deep relaxation: partial differential equations for optimizing deep neural networks.
Research in the Mathematical Sciences, 5(3):30.

[0] Cheng, X. and Bartlett, P. (2017).
Convergence of langevin mcmc in kl-divergence.
arXiv preprint arXiv:1705.09048.

[0] Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2017).
Underdamped langevin mcmc: A non-asymptotic analysis.
arXiv preprint arXiv:1707.03663.

[0] Dalalyan, A. S. and Karagulyan, A. G. (2017).
User-friendly guarantees for the langevin monte carlo with inaccurate gradient.
arXiv preprint arXiv:1710.00095.

[0] Dasgupta, P. and Maskin, E. (1986).
The existence of equilibrium in discontinuous economic games, i: Theory.
The Review of economic studies, 53(1):1–26.

[0] Durmus, A., Majewski, S., and Miasojedow, B. (2018).
Analysis of langevin monte carlo via convex optimization.
arXiv preprint arXiv:1802.09188.

[0] Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2018).
Log-concave sampling: Metropolis-hastings algorithms are fast!
arXiv preprint arXiv:1801.02309.

[0] Dziugaite, G. K. and Roy, D. (2018).

Entropy-sgd optimizes the prior of a pac-bayes bound: Generalization properties of entropy-sgd and data-dependent priors.
In *International Conference on Machine Learning*, pages 1376–1385.

[0] Fujimoto, S., van Hoof, H., and Meger, D. (2018).
Addressing function approximation error in actor-critic methods.
arXiv preprint arXiv:1802.09477.

[0] Gao, X., Gurbuzbalaban, M., and Zhu, L. (2018).
Breaking reversibility accelerates langevin dynamics for global non-convex optimization.
arXiv preprint arXiv:1812.07725.

[0] Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990).
Illustration of bayesian inference in normal data models using gibbs sampling.
Journal of the American Statistical Association, 85(412):972–985.

[0] Glicksberg, I. L. (1952).
A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points.
Proceedings of the American Mathematical Society, 3(1):170–174.

[0] Hastings, W. K. (1970).
Monte carlo sampling methods using markov chains and their applications.

[0] Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. (2018a).
Mirrored langevin dynamics.
In *Advances in Neural Information Processing Systems*.

[0] Hsieh, Y.-P., Liu, C., and Cevher, V. (2018b).
Finding mixed nash equilibria of generative adversarial networks.
arXiv preprint arXiv:1811.02002.

[0] Hsieh, Y.-P., Liu, C., and Cevher, V. (2019).
Finding mixed nash equilibria of generative adversarial networks.
In *Submitted to International Conference on Learning Representations*.
under review.

[0] Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.

[0] Kahn, H. and Harris, T. E. (1951).
Estimation of particle transmission by random sampling.
National Bureau of Standards applied mathematics series, 12:27–30.

[0] Kangarshahi*, E. A., Hsieh*, Y.-P., Sahin, M. F., and Cevher, V. (2018).
Let's be honest: An optimal no-regret framework for zero-sum games.
In *Proceedings of the 35th International Conference on Machine Learning*, pages 2488–2496.

[0] Kemeny, J. G. and Snell, J. L. (1976).
Markov Chains.
Springer-Verlag, New York.

[0] Li, C., Chen, C., Carlson, D., and Carin, L. (2016).
Preconditioned stochastic gradient langevin dynamics for deep neural networks.
In *Thirtieth AAAI Conference on Artificial Intelligence*.

[0] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015).
Continuous control with deep reinforcement learning.
arXiv preprint arXiv:1509.02971.

[0] Metropolis, N. and Ulam, S. (1949).
The monte carlo method.
Journal of the American statistical association, 44(247):335–341.

[0] Neal, R. M. et al. (2011).
Mcmc using hamiltonian dynamics.
Handbook of markov chain monte carlo, 2(11):2.

[0] Nemirovski, A. (2004).
Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems.
SIAM Journal on Optimization, 15(1):229–251.

[0] Nemirovskii, A. and Yudin, D. B. (1983).
Problem complexity and method efficiency in optimization.
Wiley.

[0] Neumann, J. v. (1928).
Zur theorie der gesellschaftsspiele.
Mathematische annalen, 100(1):295–320.

[0] Patterson, S. and Teh, Y. W. (2013).

Stochastic gradient riemannian langevin dynamics on the probability simplex.
In *Advances in Neural Information Processing Systems*, pages 3102–3110.

[0] Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen, R. Y., Chen, X., Asfour, T., Abbeel, P., and Andrychowicz, M. (2017).
Parameter space noise for exploration.
arXiv preprint arXiv:1706.01905.

[0] Sion, M. (1958).
On general minimax theorems.
Pacific Journal of mathematics, 8(1):171–176.

[0] Welling, M. and Teh, Y. W. (2011).
Bayesian learning via stochastic gradient langevin dynamics.
In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.

[0] Wibisono, A. (2018).
Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem.
arXiv preprint arXiv:1802.08089.

[0] Xu, P., Chen, J., Zou, D., and Gu, Q. (2018).
Global convergence of langevin dynamics based algorithms for nonconvex optimization.
In *Advances in Neural Information Processing Systems*, pages 3122–3133.