A connection between Tempering and Entropic Mirror Descent

Francesca R. Crucinio (King's College London)

Joint work with Nicolas Chopin and Anna Korba (CREST, ENSAE, IP Paris)

More info: arXiv preprint arXiv:2310.11914

Sampling

- **Aim 1:** sample from a probability distribution π on $\mathbb{R}^d$ and approximate expectations w.r.t. $\pi(x) = \eta(x)/\mathcal{Z}$ whose normalising constant might be unknown

$$\int f(x)\pi(x)\mathrm{d}x$$

- **Motivation:** compute posterior expectations in Bayesian inference
- **Aim 2:** estimate the unknown normalising constant $\mathcal{Z}$
- **Motivation:** model selection/parameter inference

Sampling as optimisation over distributions

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi)$$

where $\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log(\mu/\pi)\mathrm{d}\mu$ denotes the Kullback–Leibler divergence.

- Variational Inference (Blei et al., 2017)
- Algorithms based on the Langevin diffusion (Jordan et al., 1998)
- Stein Variational Gradient Descent (SVGD; Liu (2017))
- Algorithms based on tempering (this work and Domingo-Enrich and Pooladian (2023))

Gradient Descent

Gradient descent in **Euclidean space** amounts to solving

$$\dot{x}_t = -\nabla \mathcal{F}(x_t)$$

Gradient descent in the **space of distributions** amounts to solving

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \nabla \operatorname{KL}(\mu_t | \pi)\right)$$

Algorithms based on the Langevin diffusion (ULA, MALA, HMC, etc.), SVGD and continuous time tempering all implement gradient for the KL in different geometries.

Gradient Descent

Langevin diffusion $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t$

Geometry: Wasserstein-2

Gradient: $\nabla_{W_2} \mathrm{KL}(\mu_t|\pi) = \nabla \log\left(\frac{\mu_t}{\pi}\right)$

Stein Variational Gradient Descent

$dX_t^i = 1/N \sum_{j=1}^{N} \left[k(X_t^i, X_t^j)\nabla \log \pi(X_t^j) - \nabla_1 k(X_t^j, X_t^i)\right]$

Geometry: Stein

Gradient: $\nabla_{\mathrm{Stein}} \mathrm{KL}(\mu_t|\pi) = \int k(x, \cdot)\nabla \log\left(\frac{\mu_t}{\pi}(x)\right) d\mu_t(x)$

Tempering/Annealing

In the Monte Carlo literature, it is common to consider the following **tempering (or annealing)** sequence

$$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}},$$

where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T = 1$.

- Parallel Tempering (Geyer, 1991)
- Annealed Importance Sampling (Neal, 2001)
- Sequential Monte Carlo samplers (Del Moral et al., 2006)
- Termodynamic Integration (Gelman and Meng, 1998)

Why tempering?

1. can tackle multimodal targets
2. normalising constant estimated for free
3. used as alternatives to poorly mixing MCMC algorithms

Mirror Descent

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ be a functional on $\mathcal{P}(\mathbb{R}^d)$. **Mirror Descent** proceeds iteratively solving (Aubin-Frankowski et al., 2022)

$$\mu_{n+1} = \underset{\mu \in \mathcal{P}(\mathbb{R}^d)}{\operatorname{argmin}} \left\{ \mathcal{F}(\mu_n) + \langle \nabla \mathcal{F}(\mu_n), \mu - \mu_n \rangle + (\gamma_{n+1})^{-1} B_\phi(\mu | \mu_n) \right\}. \quad (1)$$

- $(\gamma_n)_{n \geq 0}$ is a sequence of step-sizes
- $B_\phi(\nu|\mu) = \phi(\nu) - \phi(\mu) - \langle \nabla \phi(\mu), \nu - \mu \rangle$ for some positive and convex ϕ is the **Bregman divergence**
- $\langle \nabla \mathcal{F}(\nu), \xi \rangle = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\nu + \epsilon \xi) - \mathcal{F}(\nu))$ is the **first variation** of $\mathcal{F}$

Entropic Mirror Descent (MD)

Using the first order conditions of (1) we obtain the dual iteration

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\gamma_{n+1}\nabla\mathcal{F}(\mu_n).$$

In the case $B_\phi(\nu|\mu) = \mathrm{KL}(\nu|\mu)$, $\nabla\phi(\mu) = \log\mu$ and we have the following multiplicative update named **entropic mirror descent**:

$$\mu_{n+1} \propto \mu_n e^{-\gamma_{n+1}\nabla\mathcal{F}(\mu_n)}.$$

Entropic Mirror Descent (MD)

Using the first order conditions of (1) we obtain the dual iteration

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\gamma_{n+1}\nabla\mathcal{F}(\mu_n).$$

In the case $B_\phi(\nu|\mu) = \mathrm{KL}(\nu|\mu)$, $\nabla\phi(\mu) = \log\mu$ and we have the following multiplicative update named **entropic mirror descent**:

$$\mu_{n+1} \propto \mu_n e^{-\gamma_{n+1}\nabla\mathcal{F}(\mu_n)}.$$

If $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$, $\nabla\mathcal{F}(\mu) = \log(\frac{\mu}{\pi})$ and we obtain entropic mirror descent on the KL:

$$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})}\pi^{\gamma_{n+1}}.$$

Connection between Tempering and MD

MD

$$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})} \pi^{\gamma_{n+1}}$$

Tempering

$$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}$$

are equivalent if

$$\lambda_n = 1 - \prod_{k=1}^{n} (1 - \gamma_k).$$

Connection between Tempering and MD

MD	Tempering
$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})} \pi^{\gamma_{n+1}}$	$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}$

are equivalent if

$$\lambda_n = 1 - \prod_{k=1}^{n} (1 - \gamma_k).$$

The connection between MD and tempering allows us to obtain explicit **rates of convergence** for the tempering iterates:

$$\mathrm{KL}(\mu_n|\pi) \leq \frac{\prod_{k=1}^{n}(1-\gamma_k)}{\gamma_1} \mathrm{KL}(\pi|\mu_0) = \frac{1-\lambda_n}{\lambda_1} \mathrm{KL}(\pi|\mu_0).$$

Choice of tempering sequence

The tempering iterates $\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}$ can be written in exponential family form

$$\mu_{n+1}(x) \equiv \mu_{\lambda_{n+1}}(x) \propto \mu_0 \exp\left\{\lambda_{n+1} s(x)\right\}$$

where $s(x) := \log \pi(x)/\mu_0(x)$.

We can compute the f-divergence between two successive iterates

$$\int \mu_\lambda f(\mu_{\lambda'}/\mu_\lambda) = \frac{f''(1)\mathrm{I}(\lambda)}{2} \times (\lambda' - \lambda)^2 + \mathcal{O}\left((\lambda' - \lambda)^3\right),$$

where $\mathrm{I}(\lambda) = \mathrm{Var}_{\mu_\lambda}\left[s(X)\right]$ is the Fisher information.

Adaptive choice of tempering

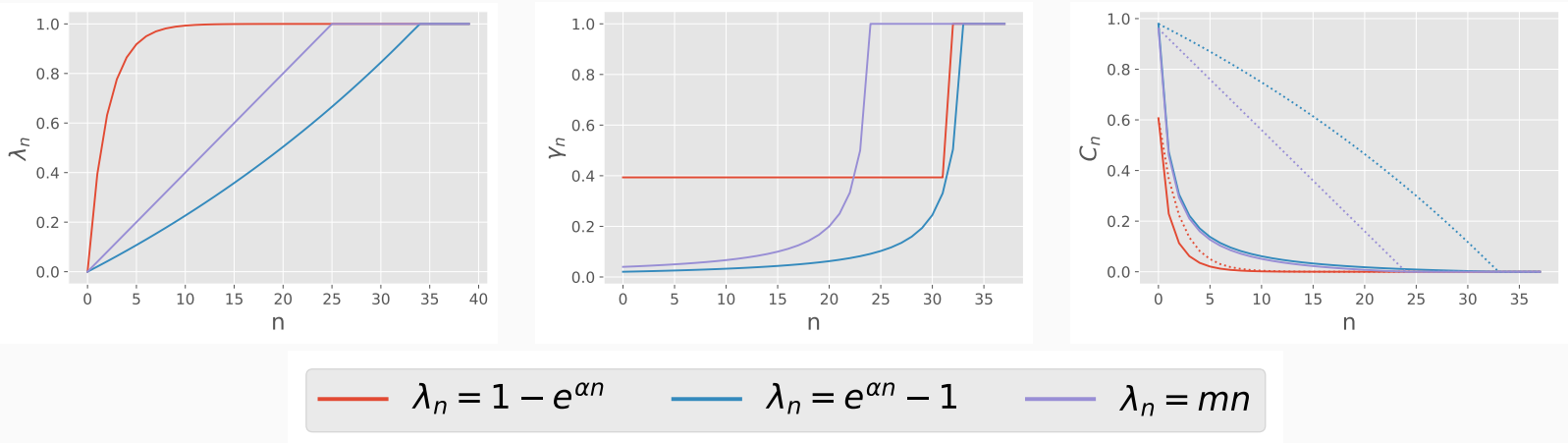
Intuitively, the distance between successive iterates should be small and constant. This suggests the following recipe to choose successive λ_n values:

$$\lambda_n - \lambda_{n-1} = c\mathrm{I}(\lambda_{n-1})^{-1/2} \tag{2}$$

Adaptive choice of tempering

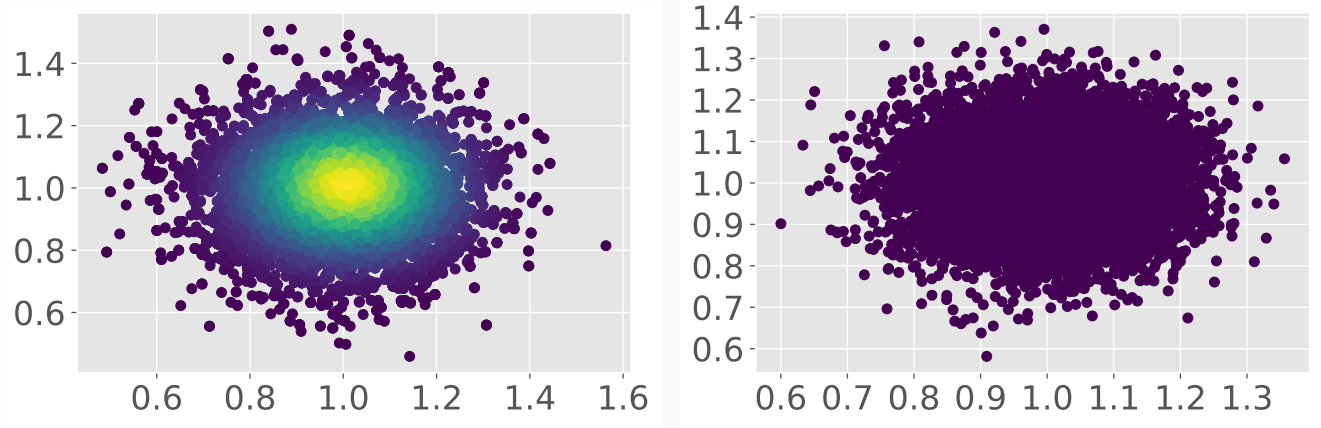
Intuitively, the distance between successive iterates should be small and constant. This suggests the following recipe to choose successive λ_n values:

$$\lambda_n - \lambda_{n-1} = c\mathrm{I}(\lambda_{n-1})^{-1/2} \tag{2}$$



Example

Approximations of $\pi = \mathcal{N}(1_d, 0.1^2 Id)$ from $\mu_0 = \mathcal{N}(0_d, Id)$.



Left: Adaptive SMC, Right: Fixed γ SMC.

Algorithms

$$\mu_{n+1} \propto q_n \exp(-\gamma_n g_n)$$

where g_n is an approximation of the gradient of the KL objective $\log(\mu_n/\pi)$; and q_n is an approximation of μ_n.

We focus on algorithms which use:

- importance weights corresponding to $\exp(-\gamma_n g_n)$
- mixtures corresponding to q_n

Sequential Monte Carlo (SMC) samplers

In SMC (Del Moral et al., 2006), the mirror descent iterate at time n is approximated by $q_n^{\mathrm{SMC}}(x) = \sum_{i=1}^N W_n^i \delta_{X_n^i}(x)$

- $\{X_n^i, W_n^i\}_{i=1}^N$ weighted particle set with

$$W_n(x) = \left(\frac{\pi(x)}{\mu_0(x)}\right)^{\lambda_n - \lambda_{n-1}} = \left(\frac{\pi(x)}{\mu_{n-1}(x)}\right)^{\gamma_n}. \tag{3}$$

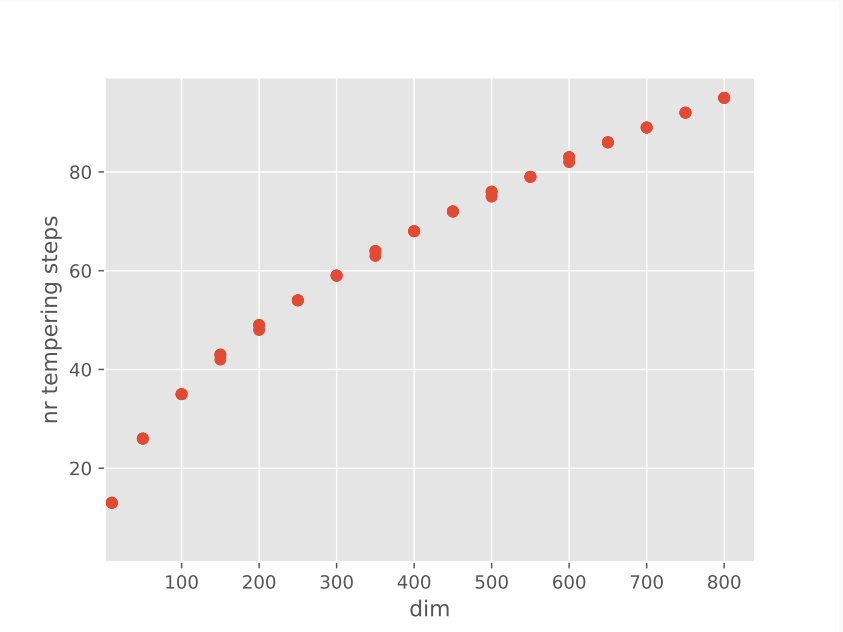
- at each iteration a new N-particle set is resampled using W_n^i and a μ_n-invariant Markov kernel

Adaptive strategies

The SMC literature offers an easy way to tune the stepsize/tempering sequence adaptively: aim for iterates which keep constant

$$\mathrm{ESS}_n(\lambda) := 1/\sum_{i=1}^{N}(W_n^i)^2.$$

1. easy and inexpensive to approximate with particle cloud
2. approximates the χ^2 divergence $\chi^2(\mu_{\lambda'}|\mu_\lambda) \approx \frac{N}{\mathrm{ESS}_n(\lambda)} - 1$
3. guarantees $T = \mathcal{O}(\sqrt{d})$



Other algorithms

1. Annealed Importance Sampling (AIS; Neal (2001))
   - same as SMC but without resampling
2. Particle Mirror Descent (PMD; Dai et al. (2016))
   - approximates μ_n with a kernel density estimator $q_n^{\mathrm{PMD}}(x) := \sum_{i=1}^{N} V_n^i K_{h_n}(x - X_n^i)$
   - uses an approximation of $\left(\frac{\pi(x)}{\mu_{n-1}(x)}\right)^{\gamma_n}$
3. Safe and Regularized Adaptive Importance Sampling (SRAIS; Korba and Portier (2022))
   - approximates μ_n with a kernel density estimator $q_n^{\mathrm{SRAIS}}(x) = \sum_{i=1}^{n} U_i K_{h_i}(x - X_i)$
   - uses an approximation of $\left(\frac{\pi(x)}{\mu_{n-1}(x)}\right)^{\gamma_n}$

Gradient flow with the Fisher-Rao geometry

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ be a functional on $\mathcal{P}(\mathbb{R}^d)$. The gradient flow of F w.r.t. the Fisher-Rao geometry

$$d_H(\nu_1, \nu_2)^2 = 4 \int (\sqrt{\nu_1} - \sqrt{\nu_2})^2$$

can be written as (Domingo-Enrich and Pooladian, 2023; Lu et al., 2023)

$$\frac{\partial \mu_t}{\partial t} = -\mu_t \mathcal{F}'(\mu_t), \text{ hence, } \frac{\partial \log(\mu_t)}{\partial t} = -\mathcal{F}'(\mu_t).$$

Mirror descent (and tempering!) can be obtained as an Euler discretisation of the FR gradient flow.

Conclusions

- the connection between mirror descent (MD) and tempering justifies tempering from an optimisation point of view and provides the MD literature with several classes of algorithms (which are very well-studied!)
- opens the door to extensions of tempering through the use of other divergences
- gives a strategy to select γ/λ adaptively

Conclusions

- the connection between mirror descent (MD) and tempering justifies tempering from an optimisation point of view and provides the MD literature with several classes of algorithms (which are very well-studied!)
- opens the door to extensions of tempering through the use of other divergences
- gives a strategy to select γ/λ adaptively

Thank you!

References

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bo Dai, Niao He, Hanjun Dai, and Le Song. Provable Bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994. PMLR, 2016.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

Carles Domingo-Enrich and Aram-Alexandre Pooladian. An Explicit Expansion of the Kullback-Leibler Divergence along its Fisher-Rao Gradient Flow. *arXiv preprint arXiv:2302.12229*, 2023.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185, 1998.

Charles J Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramides, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Anna Korba and François Portier. Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pages 11503–11527. PMLR, 2022.

Qiang Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf`.

Yulong Lu, Dejan Slepčev, and Lihan Wang. Birth–death dynamics for sampling: global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731, 2023.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

Particle Mirror Descent (PMD)

In PMD (Dai et al., 2016), the mirror descent iterate at time n is approximated by a kernel density estimator
$q_n^{\mathrm{PMD}}(x) := \sum_{i=1}^N V_n^i K_{h_n}(x - X_n^i)$

- $\{X_n^i, V_n^i\}_{i=1}^N$ weighted particle set with

$$V_n(x) = \left(\frac{\pi(x)}{q_{n-1}^{\mathrm{PMD}}(x)} \right)^{\gamma_n}.$$

- K_{h_n} a smoothing kernel with bandwidth h_n
- at each iteration a new N-particle set is resampled from q_n^{PMD}

Safe and Regularized Adaptive Importance Sampling (SRAIS)

In SRAIS (Korba and Portier, 2022), the mirror descent iterate at time n is approximated by a kernel density estimator $q_n^{\mathrm{SRAIS}}(x) = \sum_{i=1}^{n} U_i K_{h_i}(x - X_i)$

- $\{X_n^i, U_n^i\}_{i=1}^N$ weighted particle set with

$$U_n(x) = \left(\frac{\pi(x)}{q_{n-1}^{\mathrm{SRAIS}}(x)} \right)^{\gamma_n}.$$

- K_{h_i} a smoothing kernel with bandwidth h_i
- at each iteration a new particle is added sampling from q_n^{SRAIS}