The Shifted Composition Rule

Other Conference Item

Author(s):
Altschuler, Jason M.; Chewi, Sinho

Publication date:
2024-03-06

Permanent link:
https://doi.org/10.3929/ethz-b-000664602

The shifted composition rule

Jason M. Altschuler
University of Pennsylvania
email: alts@upenn.edu

Sinho Chewi
Institute for Advanced Study
email: schewi@ias.edu

This is an extended abstract for an invited talk based on [AC23; AC24].

I. Shifted composition rule

We formulate a new technique for bounding information-theoretic divergences. For KL divergence, this Shifted Chain Rule (SCR) states

$$\mathsf{KL}(\boldsymbol{\mu}^Y \,\|\, \boldsymbol{\nu}^Y) \le \mathsf{KL}(\boldsymbol{\mu}^{X'} \,\|\, \boldsymbol{\nu}^X) + \mathbb{E}\,\mathsf{KL}(\boldsymbol{\mu}^{Y|X=x} \,\|\, \boldsymbol{\nu}^{Y|X=x'})$$

where $\boldsymbol{\mu}$ is a joint distribution on X, X', Y; $\boldsymbol{\nu}$ is a joint distribution on X, Y; and the expectation is over any coupling (x, x') of $\boldsymbol{\mu}^X$ and $\boldsymbol{\mu}^{X'}$. By taking $X = X'$, the SCR generalizes the standard KL chain rule which (combined with data-processing) gives the bound

$$\begin{aligned}\mathsf{KL}(\boldsymbol{\mu}^Y \,\|\, \boldsymbol{\nu}^Y) &\le \mathsf{KL}(\boldsymbol{\mu}^{X,Y} \,\|\, \boldsymbol{\nu}^{X,Y}) \\ &= \mathsf{KL}(\boldsymbol{\mu}^X \,\|\, \boldsymbol{\nu}^X) + \mathbb{E}\,\mathsf{KL}(\boldsymbol{\mu}^{Y|X=x} \,\|\, \boldsymbol{\nu}^{Y|X=x})\,.\end{aligned}$$

The key advantage of the SCR is the additional flexibility in X', which intuitively enables modifying the "history" of the process $X \mapsto Y$ to $X' \mapsto Y$ (first term) at a price given by how different X and X' are (second term). This enables addressing applications where the standard chain rule would not suffice, such as situations where $\mathsf{KL}(\boldsymbol{\mu}^X \,\|\, \boldsymbol{\nu}^X)$ is large or even infinite (e.g., $\boldsymbol{\mu}^X$, $\boldsymbol{\nu}^X$ are different Dirac measures).

More generally, our papers consider Rényi divergences of any positive order. The SCR then becomes the Shifted Composition Rule, analogously extending the standard Rényi composition rule via this additional flexibility in X'. In this abstract, we focus on KL for simplicity of exposition.

II. Reverse transport inequalities

In these two papers, our main application is the derivation of reverse transport inequalities for the Langevin diffusion

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{1}$$

where $(B_t)_{t\ge 0}$ is Brownian motion. Let $(P_t)_{t\ge 0}$ denote the Markov semigroup for (1) and fix probability measures μ, ν, ν' and $t > 0$. As a representative example of our results, we use the SCR to show that if $\nabla^2 V \succeq \alpha I$, then

$$\mathsf{KL}(\mu P_t \,\|\, \nu P_t) \le \frac{\alpha\, W_2^2(\mu,\nu)}{2\,(\exp(2\alpha T) - 1)}\,, \tag{2}$$

and if $-\beta I \preceq \nabla^2 V \preceq \beta I$, then

$$\mathsf{KL}(\mu P_t * \nu \,\|\, \mu P_t * \nu') \le \frac{\beta\, W_2^2(\nu,\nu')}{2\,(1 - \exp(-2\beta t))}\,. \tag{3}$$

These inequalities capture complementary aspects of the diffusion: (2) measures sensitivity w.r.t. the initial condition (indeed, for $\alpha \ge 0$ it yields a mixing time bound), whereas (3) captures the regularity of the marginal law of the process. In other words, they encode regularity for Kolmogorov's backward and forward equations, respectively.

As an illustration of the use of the SCR, suppose that we want to establish (2) when $\mu = \delta_x$, $\nu = \delta_y$ are Diracs. To formulate an argument in discrete time, we first replace the continuous-time semigroup $(P_t)_{t\ge 0}$ by a discretized one and apply a limiting argument. Then, the question is to bound $\mathsf{KL}(\delta_x P^N \,\|\, \delta_y P^N)$ for a Markov kernel P. A naïve application of the KL chain rule is vacuous, since $\mathsf{KL}(\delta_x \,\|\, \delta_y) = \infty$. Instead, we construct an auxiliary process $\{X'_n\}_{n=0}^N$ such that $X'_0 = y$ and $X'_N \sim \delta_x P^N$, and we instead bound $\mathsf{KL}(\mathrm{law}(X'_N) \,\|\, \delta_y P^N)$ via the SCR (details in [AC23]). In this context, this argument can be seen as a generalization of the shifted divergence technique from the differential privacy and sampling literature [Fel+18; AT22; AT23] or as a discrete-time analogue of the coupling in [ATW06].

III. Functional analysis, geometry, and probability

Inequalities (2) and (3) are part of a larger story—called *Bakry–Émery* theory [BGL14]—which relates analytic properties of the semigroup, through functional inequalities, to the curvature of the underlying space and of the measure (i.e., the Hessian $\nabla^2 V$ of the negative log-density), and to probabilistic aspects such as concentration of measure and mixing. Indeed, it is well-known that via duality, (2) is equivalent to the celebrated dimension-free *Harnack* inequality of [Wan97], and implies back the curvature lower bound $\nabla^2 V \succeq \alpha I$.

On the other hand, the inequality (3), which is equivalent to a *shift Harnack* inequality [Wan14], appears in its sharp form for the first time in our paper [AC24]. This allows us to prove that (3) implies back the curvature *upper bound* $\nabla^2 V \preceq \beta I$. In our paper, we leave open the intriguing question of whether this observation can form the basis of a Bakry–Émery theory for curvature upper bounds.

References

[AC24] J. M. Altschuler and S. Chewi. "Shifted composition II: shift Harnack inequalities and curvature upper bounds". In: *arXiv preprint arXiv:2401.00071* (2024).

[AC23] J. M. Altschuler and S. Chewi. "Shifted composition I: Harnack and reverse transport inequalities". In: *arXiv preprint 2311.14520* (2023).

[AT22] J. M. Altschuler and K. Talwar. "Privacy of noisy stochastic gradient descent: more iterations without more privacy loss". In: *Advances in Neural Information Processing Systems*. 2022.

[AT23] J. M. Altschuler and K. Talwar. "Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling". In: *Conference on Learning Theory*. 2023.

[ATW06] M. Arnaudon, A. Thalmaier, and F.-Y. Wang. "Harnack inequality and heat kernel estimates on manifolds with curvature unbounded below". In: *Bull. Sci. Math.* 130.3 (2006), pp. 223–233.

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[Fel+18] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. "Privacy amplification by iteration". In: *Symposium on Foundations of Computer Science*. 2018.

[Wan97] F.-Y. Wang. "Logarithmic Sobolev inequalities on noncompact Riemannian manifolds". In: *Probability Theory and Related Fields* 109.3 (1997), pp. 417–424.

[Wan14] F.-Y. Wang. "Integration by parts formula and shift Harnack inequality for stochastic equations". In: *The Annals of Probability* 42.3 (2014), pp. 994–1019.