Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks *

Pratik Chaudhari and Stefano Soatto
Computer Science, University of California, Los Angeles
Email: pratikac@ucla.edu, soatto@ucla.edu

Abstract

Stochastic gradient descent (SGD) is widely believed to perform implicit regularization, but the precise manner in which this occurs has thus far been elusive. We prove that SGD minimizes an average potential over the posterior distribution of the weights along with an entropic regularization term. This potential is however not the original loss function in general; the two are equivalent if and only if gradient noise in SGD is isotropic. We show that the covariance matrix of mini-batch gradients for deep networks is highly non-isotropic, with rank as small as 1% of its dimension. Moreover, SGD does not converge in the classical sense; due to non-isotropic gradient noise, most likely trajectories of SGD are closed loops in the weight space instead of being Brownian motion around critical points.

1 Introduction

Our first result shows that for a loss function $f(x)$ with weights $x \in \Omega \subset \mathbb{R}^d$, if $\rho^{\text{ss}}(x)$ is the steady-state posterior distribution over the weights as estimated by SGD,

$$\rho^{\text{ss}} = \arg\min_{\rho} \; \mathbb{E}_{x\sim\rho}\Big[\Phi(x)\Big] - \frac{\eta}{2\mathscr{b}}\, H(\rho).$$

Here, $H(\rho)$ is the entropy of the distribution ρ and η and $\mathscr{b}$ are the learning rate and batch-size, respectively. The implicit potential $\Phi(x)$ is only a function of the architecxwture and the dataset. This implies that SGD implicitly performs variational inference with a uniform prior, albeit using a different loss than the one used to compute gradients.

The potential $\Phi(x)$ is equal to our chosen loss $f(x)$ if and only if the noise due to mini-batch gradients is isotropic. This condition, however, is *not* satisfied for deep networks. Empirically, we find gradient noise to be highly non-isotropic with the rank of its covariance matrix being about 1% of its dimension. Most likely locations of SGD which are the critical points of $\Phi(x)$ can be significantly different from critical points of $f(x)$; the deviation scales linearly with $\eta/\mathscr{b}$. Furthermore, most likely trajectories of SGD are "limit cycles", i.e., closed loops in the weight space, instead of being Brownian motion around critical points. This is our second main result.

We discuss practical implications of these results for training with large batch-sizes, Bayesian inference along with hyper-parameter and neural architecture search.

2 Background on continuous-time SGD

Stochastic gradient descent performs the updates $x_{k+1} = x_k - \eta\, \nabla f_{\mathscr{b}}(x_k)$ while training a network where η is the learning rate and $\nabla f_{\mathscr{b}}(x_k)$ is the average back-propagation gradient over a mini-batch of size $\mathscr{b}$. We assume that weights belong to a compact subset $\Omega \subset \mathbb{R}^d$ to ensure appropriate boundary conditions for the evolution of steady-state densities in SGD, although all our results hold without this assumption if the loss grows unbounded as $\|x\| \to \infty$; for instance, with weight decay as a regularizer.

*This article summarizes the findings in [1]. See the longer version for background and detailed results.

Presented at the Workshop on Advances in Approximate Bayesian Inference (NIPS 2017).

Definition 1 (**Diffusion matrix** $D(x)$)**.** If examples in a mini-batch are sampled with replacement, the variance of mini-batch gradients is $\text{var}\Big(\nabla f_{\mathscr{b}}(x)\Big) = \frac{D(x)}{\mathscr{b}}$ where

$$D(x) = \left(\frac{1}{N}\sum_{k=1}^{N} \nabla f_k(x)\, \nabla f_k(x)^\top\right) - \nabla f(x)\, \nabla f(x)^\top \succeq 0. \tag{1}$$

Note that $D(x)$ is independent of the learning rate η and the batch-size $\mathscr{b}$. It only depends on the weights x, architecture and loss defined by $f(x)$, and the dataset.

Lemma 2 (**Continuous-time SGD**)**.** *The continuous-time limit of SGD is given by*

$$dx(t) = -\nabla f(x)\, dt + \sqrt{2\beta^{-1}\, D(x)}\, dW(t); \tag{2}$$

where $W(t)$ is Brownian motion and the parameter β is the inverse temperature defined as $\beta^{-1} = \frac{\eta}{2\mathscr{b}}$. The steady-state density of the weights $\rho(x,t) \propto \mathbb{P}\big(x(t) = x \,|\, x(0)\big)$, evolves according to the Fokker-Planck equation [2]:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \Big(\nabla f(x)\, \rho + \beta^{-1} D(x)\, \nabla \rho\Big) \tag{FP}$$

where the notation $\nabla \cdot v$ denotes the divergence $\nabla \cdot v = \sum_i \partial_{x_i} v_i(x)$ for any vector $v(x) \in \mathbb{R}^d$.

Note that β^{-1} completely captures the magnitude of noise in SGD that depends on the learning rate η and the mini-batch size $\mathscr{b}$.

3 SGD performs variational inference

We assume that the steady-state distribution of the Fokker-Planck equation (FP) exists, this is denoted by $\rho^{\text{ss}}(x)$ and satisfies $\rho_t^{\text{ss}} = 0$. We use this solution to implicitly define a potential

$$\Phi(x) = -\beta^{-1} \log \rho^{\text{ss}}(x), \tag{3}$$

up to a constant. This allows us to write $\rho^{\text{ss}}(x) = Z^{-1}(\beta)\; e^{-\beta\Phi(x)}$ where $Z(\beta)$ is a normalizing constant. The potential will be made explicit in Section 4.

It can be shown by direct substitution in (FP) that this same distribution is also the steady-state solution of (2) if $\nabla f(x) = D\, \nabla\Phi(x)$. We exploit this to split the gradient $\nabla f(x)$ into two quantities:

$$j(x) = -\nabla f(x) + D(x)\, \nabla\Phi(x), \tag{4}$$

where $j(x)$ is the part of the gradient that cannot be written as the gradient of some potential Φ'. We now make an important assumption on $j(x)$ which has its origins in thermodynamics.

Assumption 3 (**Force** $j(x)$ **is conservative**)**.** We assume that $\nabla \cdot j(x) = 0$.

This assumption is discussed further in Appendix B and an important implication is that $j(x)$ is orthogonal to $\nabla\Phi(x)$. Our first main result is the following theorem.

Theorem 4 (**SGD performs variational inference**)**.** *The functional*

$$F(\rho) = \beta^{-1}\, \text{KL}\big(\rho \,||\, \rho^{\text{ss}}\big) \tag{5}$$

decreases monotonically along the trajectories of the Fokker-Planck equation (FP) and converges to its minimum, which is zero, at steady-state. Moreover, we also have an energetic-entropic split

$$F(\rho) = \mathbb{E}_{x\in\rho}\, \Big[\Phi(x)\Big] - \beta^{-1} H(\rho) + \text{constant}. \tag{6}$$

Theorem 4 shows that SGD implicitly minimizes a combination of two terms. The "energetic" term is the average potential over the distribution ρ and encourages ρ^{ss} to place its probability mass in regions of the parameter space with small values of Φ. The "entropic" term is a regularization term that biases SGD towards solutions that maximize the entropy of the distribution ρ. Note that the energetic term in (6) has potential $\Phi(x)$, instead of $f(x)$.

Let us note that (6) becomes the celebrated JKO functional in optimal transport [3] if diffusion is isotropic. In this case, SGD performs gradient descent on (6) in the Wasserstein metric [4]. Theorem 4

and the JKO functionals provide a way to enforce desired properties on the steady-state distribution of SGD by modifying and interpreting (6). This has had enormous impact in optimal transport [5].

Lemma 5 (Potential equals original loss iff isotropic diffusion). *If the diffusion matrix $D(x)$ is isotropic, i.e., a constant multiple of the identity, the implicit potential is the original loss itself:*

$$D(x) = c\, I_{d\times d} \quad \Leftrightarrow \quad \Phi(x) = f(x). \tag{7}$$

Lemma 6 (SGD converges to limit cycles). *The force $j(x)$ does not decrease $F(\rho)$ in (6) and introduces a deterministic component in the Fokker-Planck equation (FP) given by $\dot{x} = j(x)$. Together with Assumption 3, this implies that most likely trajectories of SGD are closed loops.*

Remark 7 (Connection to Bayesian inference). Note that (6) corresponds to the evidence lower bound [6] with a uniform prior on Ω with two differences: (i) the multiplier of β, and (ii) the energetic term has the potential $\Phi(x)$ instead of the loss $f(x)$ as is usually in the Bayesian literature.

Remark 8 (ELBO in practice). In practice, optimizing ELBO involves one or multiple steps of SGD to minimize the energetic term along with an analytically computable KL-divergence term which is often achieved using a factored Gaussian prior [6]. As Theorem 4 shows, such an approach implicitly also enforces a uniform prior whose strength is determined by β^{-1} and conflicts with the externally imposed Gaussian prior in ELBO. This conflict, which fundamentally arises from using SGD to minimize the energetic term, has resulted in researchers modulating the strength of the Gaussian prior in ELBO using a scalar factor in the KL-divergence term [7]. This has also been shown to lead to invariant representations [8] via the information bottleneck principle [9].

3.1 Practical implications

We show in Section 4 that the potential $\Phi(x)$ does not depend on β, it is only a function of the dataset and the architecture. Therefore, the two parameters, η and $\mathscr{b}$, completely determine the strength of the entropic regularization term. If $\beta^{-1} \to 0$, the implicit regularization of SGD goes to zero. This suggests that $\beta^{-1} = \frac{\eta}{2\mathscr{b}}$ should not be small, that learning rate should scale linearly with mini-batch size. This theoretical prediction fits very well with empirical evidence wherein one obtains good generalization performance only with small mini-batches [10], or via such linear scaling [11].

If mini-batches are sampled with replacement, the diffusion matrix is very similar, but the inverse temperature becomes $\beta'^{-1} = \frac{\eta}{2\mathscr{b}}\left(1 - \frac{\mathscr{b}}{N}\right)$. The extra factor goes to zero as $\mathscr{b} \to N$ and consequently, Theorem 4 predicts that for the same learning rate and batch-size, sampling with replacement results in improved generalization than without. This effect is particularly pronounced at large batch-sizes.

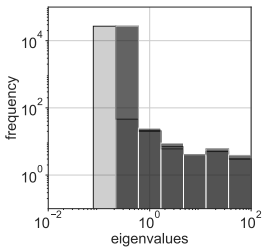

Figure 1: CIFAR-10: $\lambda(D) = 0.27 \pm 0.84$, $\text{rank}(D) = 0.34\%$

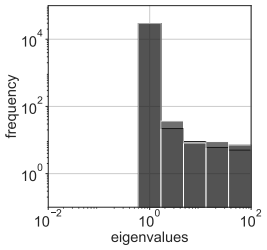

Figure 2: CIFAR-100: $\lambda(D) = 0.98 \pm 2.16$, $\text{rank}(D) = 0.47\%$

4 Characterizing the potential $\Phi(x)$

Figs. 1 and 2 show the eigenspectrum of $D(x)$ for a convolutional network on the CIFAR-10/100 datasets [12]. It has a large fraction of almost-zero eigenvalues with a very small rank that ranges between 0.3% - 2% of the dimension. The non-zero eigenvalues are spread across a large range. Also note that the eigenspectra at three different instants (20%, 40% and 100% training completion, darker is later) are almost indistinguishable. This suggests that the spectral properties of the gradient noise in deep networks are almost constant, although the matrix is highly non-isotropic. Using Lemma 5, this shows that $\Phi(x) \neq f(x)$ for deep networks.

The eigenspectrum in Fig. 2 has a larger mean and variance than the one in Fig. 1; this is expected to a more diverse dataset. Since the noise in SGD has variance $\beta^{-1}D(x)$, we see that maintaining $\beta^{-1}\text{mean}(\lambda(D))$ constant across datasets is a good way to pick hyper-parameters; this keeps the magnitude of the noise constant. Note that data-augmentation, which generalizes better, has a larger variance in the eigenspectrum. This indicates that the quantity $\frac{\text{rank}(D)}{d} + \text{var}(\lambda(D))$ can be used for automated neural architecture search, possibly even without training the network.

4.1 Explicit formula for Φ

The Ornstein-Uhlenbeck equation [13] gives a closed-form solution for the case when ∇f is linearized around a critical point [7, 14]. The linear case satisfies

$$\nabla f(x) = (D+Q)\,\nabla\Phi(x) \quad \text{and} \quad j(x) = -Q\,\nabla\Phi(x), \tag{8}$$

where D and Q are symmetric and anti-symmetric parts of a unique matrix G with $G\,\nabla^2 f(0)^\top = \nabla^2 f(0)\,G^\top$. This suggests that the effect of a non-zero $j(x)$ is to rotate the local gradient field. We build upon this and posit that the general solution is

$$\nabla f(x) = \Big(D(x)+Q(x)\Big)\,\nabla\Phi(x) - j_{\text{off}}(x). \tag{9}$$

where the matrices now depend on the weights x and we have added a correction force $j_{\text{off}}(x)$.

Assumption 9 (Same steady-state distribution). We now assume that the force $j_{\text{off}}(x)$ does not change the steady-state distribution, $\rho^{\text{ss}} \propto e^{-\beta\Phi}$. This is a popular assumption in physics and biology [15–17] and effectively amounts to assuming that the symmetric matrix $D(x)$ and an anti-symmetric matrix $Q(x)$ captures most of the gradient $\nabla f(x)$ that results from a potential $\Phi(x)$.

Our second main result now follows.

Theorem 10 (Most likely locations are not the critical points of the loss). *The most likely locations of SGD where $\nabla\Phi(x) = 0$ are different from the critical points of the original loss $\nabla f(x) = 0$:*

$$j_{\text{off}}(x) = \beta^{-1}\,\Big(\nabla\cdot Q(x)\Big)^\top, \tag{10}$$

where the divergence operator is applied to the matrix $Q(x)$ column-wise. The matrix $Q(x)$ and the potential $\Phi(x)$ can be explicitly computed in terms of the gradient $\nabla f(x)$ and the diffusion matrix $D(x)$. In particular, $\Phi(x)$ does not depend on β.

The time spent by a Markov chain at a state x is proportional to $\rho^{\text{ss}}(x)$. While it is easily seen that SGD does not converge in the Cauchy sense due to the stochasticity, it is very surprising that it may spend a significant amount of time away from the critical points of the original loss if $Q(x)$ has a large divergence. The proof of Theorem 10 also shows that the presence of a $Q(x)$ with non-zero divergence is the consequence of a non-isotropic $D(x)$ and it persists even if D is constant and independent of weights. This is indeed the case for deep networks from our experiments.

The effect predicted by (10) becomes more pronounced if $\beta^{-1} = \frac{\eta}{2\hat{\beta}}$ is large, i.e., for small batch-sizes or high learning rates. Theorem 4 also shows that as $\beta^{-1} \to 0$, the implicit entropic regularization in SGD vanishes. Observe that these are exactly the conditions under which we typically obtain good generalization performance for deep networks [10, 11]. This suggests that a non-isotropic gradient noise is crucial to obtain good generalization performance and indicates a link with other non-equilibrium sampling techniques used in deep learning [18, 19].

5 Discussion

The idea that SGD is related to variational inference has been seen in machine learning before [7, 20] under assumptions such as quadratic potentials; for instance, see [21] for methods to approximate steady-states by interpreting SGD as Langevin dynamics. Our results here are however different, we aim to characterize steady-state distributions of SGD itself. In full generality, SGD performs variational inference, albeit using a new potential Φ instead of the original loss $f(x)$.

Our results make precise, the widely held belief that SGD is an implicit regularizer [22–24]. SGD introduces an explicit entropic regularization on the posterior. Further, non-isotropic gradient noise for deep networks reveals a potential Φ with properties that lead to both generalization and acceleration.

Noise is often added in SGD to improve its behavior around saddle points for non-convex losses, see [25–27]. It is also quite indispensable for training deep networks [28–32]. There is however a disconnect between these two directions due to the fact that while adding external gradient noise helps in theory, it works poorly in practice [33, 34]. Instead, "noise tied to the architecture" works better, e.g., dropout, or small mini-batches. Our results close this gap and show that SGD crucially leverages the highly degenerate noise induced by the architecture.

References

[1] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *arXiv:1710.11029*, 2017.

[2] Hannes Risken. The Fokker-Planck Equation. Springer, 1996.

[3] Richard Jordan, David Kinderlehrer, and Felix Otto. Free energy and the fokker-planck equation. *Physica D: Nonlinear Phenomena*, 107(2-4):265–271, 1997.

[4] Filippo Santambrogio. Euclidean, metric, and Wasserstein gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[5] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2013.

[7] Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In *ICML*, pages 354–363, 2016.

[8] Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *arXiv:1706.01350*, 2017.

[9] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[10] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv:1609.04836*, 2016.

[11] Priya Goyal, Piotr Dollr, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv:1706.02677*, 2017.

[12] A. Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Computer Science, University of Toronto, 2009.

[13] Bernt Oksendal. Stochastic differential equations. Springer, 2003.

[14] Chulan Kwon, Ping Ao, and David J Thouless. Structure of stochastic dynamics near fixed points. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13029–13033, 2005.

[15] Ping Ao, Chulan Kwon, and Hong Qian. On the existence of potential landscape in the evolution of complex systems. *Complexity*, 12(4):19–27, 2007.

[16] Chulan Kwon and Ping Ao. Nonequilibrium steady state of a stochastic system driven by a nonlinear drift force. *Physical Review E*, 84(6):061106, 2011.

[17] Hong Qian. Thermodynamics of the general diffusion process: Equilibrium supercurrent and nonequilibrium driven circulation with dissipation. *The European Physical Journal Special Topics*, 224(5):781–799, 2015.

[18] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Lucibello Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Physical review letters*, 115(12):128101, 2015.

[19] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: biasing gradient descent into wide valleys. *arXiv:1611.01838*, 2016.

[20] David Duvenaud, Dougal Maclaurin, and Ryan Adams. Early stopping as non-parametric variational inference. In *AISTATS*, pages 1070–1077, 2016.

[21] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv:1704.04289*, 2017.

[22] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.

[23] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv:1705.03071*, 2017.

[24] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

[25] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *COLT*, pages 1246–1257, 2016.

[26] Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *COLT*, pages 81–102, 2016.

[27] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

[28] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM, 1993.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[30] Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *NIPS*, pages 2575–2583, 2015.

[31] Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. In *ICML*, pages 3059–3068, 2016.

[32] Alessandro Achille and Stefano Soatto. Information dropout: learning optimal representations through noise. *arXiv:1611.01353*, 2016.

[33] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv:1511.06807*, 2015.

[34] Pratik Chaudhari and Stefano Soatto. On the energy landscape of deep networks. *arXiv:1511.06485*, 2015.

[35] Chris Junchi Li, Lei Li, Junyang Qian, and Jian-Guo Liu. Batch size matters: A diffusion approximation framework on nonconvex stochastic gradient descent. *arXiv:1705.07562*, 2017.

[36] Hans Ottinger. *Beyond equilibrium thermodynamics*. John Wiley & Sons, 2005.

[37] Hong Qian. The zeroth law of thermodynamics and volume-preserving conservative system in equilibrium with stochastic damping. *Physics Letters A*, 378(7):609–616, 2014.

[38] Ilya Prigogine. *Thermodynamics of irreversible processes*, volume 404. Thomas, 1955.

[39] Lars Onsager. Reciprocal relations in irreversible processes. I. *Physical review*, 37(4):405, 1931.

[40] Lars Onsager. Reciprocal relations in irreversible processes. II. *Physical review*, 38(12):2265, 1931.

[41] Till Daniel Frank. *Nonlinear Fokker-Planck equations: fundamentals and applications*. Springer Science & Business Media, 2005.

[42] Edwin T Jaynes. The minimum entropy production principle. *Annual Review of Physical Chemistry*, 31(1):579–601, 1980.

A Diffusion matrix $D(x)$

In this section we denote $g_k := \nabla f_k(x)$ and $g := \nabla f(x) = \frac{1}{N}\sum_{k=1}^{N} g_k$. Although we drop the dependence of g_k on x to keep the notation clear, we emphasize that the diffusion matrix D depends on the weights x.

A.1 With replacement

Let $i_1, \ldots, i_{\mathscr{b}}$ be $\mathscr{b}$ iid random variables in $\{1, 2, \ldots, N\}$. We would like to compute

$$\mathrm{var}\left(\frac{1}{\mathscr{b}}\sum_{j=1}^{\mathscr{b}} g_{i_j}\right) = \mathbb{E}_{i_1,\ldots,i_{\mathscr{b}}}\left\{\left(\frac{1}{\mathscr{b}}\sum_{j=1}^{\mathscr{b}} g_{i_j} - g\right)\left(\frac{1}{\mathscr{b}}\sum_{j=1}^{\mathscr{b}} g_{i_j} - g\right)^{\top}\right\}.$$

Note that we have that for any $j \neq k$, the random vectors g_{i_j} and g_{i_k} are independent. We therefore have

$$\mathrm{covar}(g_{i_j}, g_{i_k}) = 0 = \mathbb{E}_{i_j,\, i_k}\left\{(g_{i_j} - g)(g_{i_k} - g)^{\top}\right\}$$

We use this to obtain

$$\begin{aligned}\mathrm{var}\left(\frac{1}{\mathscr{b}}\sum_{j=1}^{\mathscr{b}} g_{i_j}\right) = \frac{1}{\mathscr{b}^2}\sum_{j=1}^{\mathscr{b}} \mathrm{var}(g_{i_j}) &= \frac{1}{N\mathscr{b}}\sum_{k=1}^{N}\left((g_k - g)\,(g_k - g)^{\top}\right) \\ &= \frac{1}{\mathscr{b}}\left(\frac{\sum_{k=1}^{N} g_k\, g_k^{\top}}{N} - g\, g^{\top}\right).\end{aligned}$$

We will set

$$D(x) = \frac{1}{N}\left(\sum_{k=1}^{N} g_k\, g_k^\top\right) - g\, g^\top. \tag{A1}$$

and assimilate the factor of $\mathscr{b}^{-1}$ in the inverse temperature β.

A.2 Without replacement

Let us define an indicator random variable $\mathbf{1}_{i\in\mathscr{b}}$ that denotes if an example i was sampled in batch $\mathscr{b}$. We can show that

$$\text{var}(\mathbf{1}_{i\in\mathscr{b}}) = \frac{\mathscr{b}}{N} - \frac{\mathscr{b}^2}{N^2},$$

and for $i \neq j$,

$$\text{covar}(\mathbf{1}_{i\in\mathscr{b}}, \mathbf{1}_{j\in\mathscr{b}}) = -\frac{\mathscr{b}(N-\mathscr{b})}{N^2(N-1)}.$$

Similar to [35], we can now compute

$$\begin{aligned}
\text{var}\left(\frac{1}{\mathscr{b}}\sum_{k=1}^{N} g_k\, \mathbf{1}_{k\in\mathscr{b}}\right) &= \frac{1}{\mathscr{b}^2}\,\text{var}\left(\sum_{k=1}^{N} g_k\, \mathbf{1}_{k\in\mathscr{b}}\right) \\
&= \frac{1}{\mathscr{b}^2}\sum_{k=1}^{N} g_k\, g_k^\top\,\text{var}(\mathbf{1}_{k\in\mathscr{b}}) + \frac{1}{\mathscr{b}^2}\sum_{i,j=1,\, i\neq j}^{N} g_i\, g_j^\top\,\text{covar}(\mathbf{1}_{i\in\mathscr{b}}, \mathbf{1}_{j\in\mathscr{b}}) \\
&= \frac{1}{\mathscr{b}}\left(1-\frac{\mathscr{b}}{N}\right)\left[\frac{\sum_{k=1}^{N} g_k\, g_k^\top}{N-1} - \left(1-\frac{1}{N-1}\right) g\, g^\top\right].
\end{aligned}$$

We will again set

$$D(x) = \frac{1}{N-1}\left(\sum_{k=1}^{N} g_k\, g_k^\top\right) - \left(1-\frac{1}{N-1}\right) g\, g^\top \tag{A2}$$

and assimilate the factor of $\mathscr{b}^{-1}\left(1-\frac{\mathscr{b}}{N}\right)$ that depends on the batch-size in the inverse temperature β.

B Discussion on Assumption 3

The Fokker-Planck equation (FP) typically models a physical system which exchanges energy with an external environment [36, 37]. In our case, this physical system is the gradient descent part $\nabla\cdot(\nabla f\,\rho)$ while the interaction with the environment is the stochastic part $\beta^{-1}\nabla\cdot(D\,\nabla\rho)$. The second law of thermodynamics states that the entropy of a system can never decrease and we now show how this assumption is sufficient to guarantee this.

Let $F(\rho)$ be as defined in (6). In non-equilibrium thermodynamics, it is assumed that if the system is not far from equilibrium, i.e., if $\|j(x)\|$ is small, the local entropy production is a product of the force $-\nabla\left(\frac{\delta F}{\delta\rho}\right)$ from (6) and the probability current $-J(x,t)$ from (FP). This assumption in this form was first introduced by [38] based on the works of Onsager [39, 40]. See [41, Sec. 4.5] for a mathematical treatment and [42] for further discussion. The rate of entropy (S_i) increase is given by

$$\beta^{-1}\,\frac{dS_i}{dt} = \int_{x\in\Omega} \nabla\left(\frac{\delta F}{\delta\rho}\right) J(x,t)\,dx.$$

This can now be written using **??** again as

$$\beta^{-1}\,\frac{dS_i}{dt} = \int \rho\, D : \left(\nabla\frac{\delta F}{\delta\rho}\right)\left(\nabla\frac{\delta F}{\delta\rho}\right)^\top + \int j\rho\left(\nabla\frac{\delta F}{\delta\rho}\right) dx.$$

The first term in the above expression is non-negative, in order to ensure that $\frac{dS_i}{dt} \geq 0$, we require

$$\begin{aligned} 0 &= \int j\rho \left(\nabla \frac{\delta F}{\delta \rho} \right) dx \\ &= \int \nabla \cdot (j\rho) \left(\frac{\delta F}{\delta \rho} \right) dx; \end{aligned}$$

where the second equality again follows by integration by parts. It can be shown [41, Sec. 4.5.5] that the condition in Assumption 3, viz., $\nabla \cdot j(x) = 0$, is sufficient to make the above integral vanish and therefore for the entropy generation to be non-negative.

We also discuss some properties of $j(x)$ in Appendix C that are a consequence of this assumption.

C Some properties of the force $j(x)$

The Fokker-Planck equation (FP) can be written in terms of the probability current as

$$\begin{aligned} \rho_t^{\text{ss}} &= \nabla \cdot \left(-j\, \rho^{\text{ss}} + D\, \nabla \Phi\, \rho^{\text{ss}} + \beta^{-1} D\, \nabla \rho^{\text{ss}} \right) \\ &= \nabla \cdot\, J^{\text{ss}}. \end{aligned}$$

Since we have $\rho^{\text{ss}} \propto e^{-\beta \Phi(x)}$, we also have that

$$0 = \rho_t^{\text{ss}} = \nabla \cdot \left(D\, \nabla \Phi\, \rho^{\text{ss}} + \beta^{-1} D\, \nabla \rho^{\text{ss}} \right),$$

and consequently,

$$\begin{aligned} 0 &= \nabla \cdot (j\, \rho^{\text{ss}}) \\ \Rightarrow \quad j(x) &= \frac{J^{\text{ss}}(x)}{\rho^{\text{ss}}(x)}. \end{aligned} \tag{A3}$$

In other words, the conservative force is non-zero only if detailed balance is broken, i.e., $J^{\text{ss}} \neq 0$. We also have

$$\begin{aligned} 0 &= \nabla \cdot (j\, \rho^{\text{ss}}) \\ &= \rho^{\text{ss}} \left(\nabla \cdot j \;-\; j \cdot \nabla \Phi \right), \end{aligned}$$

which shows using Assumption 3 and $\rho^{\text{ss}}(x) > 0$ for all $x \in \Omega$ that $j(x)$ is always orthogonal to the gradient of the potential

$$\begin{aligned} 0 &= j(x) \cdot \nabla \Phi(x) \\ &= j(x) \cdot \nabla \rho^{\text{ss}}. \end{aligned} \tag{A4}$$

Using the definition of $j(x)$ in (4), we have detailed balance when

$$\nabla f(x) = D(x)\, \nabla \Phi(x). \tag{A5}$$