Near-Optimality of Contrastive Divergence Algorithms

Pierre Glaser **Kevin Han Huang** **Arthur Gretton**
Gatsby Computational Neuroscience Unit, University College London
pierreglaser@gmail.com, han.huang.20@ucl.ac.uk, arthur.gretton@gmail.com

Abstract

We perform a non-asymptotic analysis of the contrastive divergence (CD) algorithm, a training method for unnormalized models. While prior work has established that (for exponential family distributions) the CD iterates asymptotically converge at an $O(n^{-1/3})$ rate to the true parameter of the data distribution, we show, under some regularity assumptions, that CD can achieve the parametric rate $O(n^{-1/2})$. Our analysis provides results for various data batching schemes, including the fully online and minibatch ones. We additionally show that CD can be near-optimal, in the sense that its asymptotic variance is close to the Cramér-Rao lower bound.

1 Introduction

Describing data using probability distributions is a central task in multiple scientific and industrial disciplines [1, 2, 3]. Since the true distribution of the data is generally unknown, such a task requires finding an estimator of the true distribution among a model class that best describes the available data. An estimator can be characterized at multiple levels of granularity: at the highest level lies *consistency* [4], a property which states that as the number of available data points increases, a given estimator will converge to the one best describing the data distribution. At a lower level, a consistent estimator can be further characterized by its convergence rate, a quantity upper–bounding its distance to the true distribution as a function of the number of samples. A convergence rate can be either *asymptotic*, e.g. hold only in the limit of an infinite sample size, or *non-asymptotic*, in which case the rate also holds for finite sample sizes. In their simplest form, convergence rates are provided in big–O notation, discarding finer grained information such as asymptotically dominated quantities as well as multiplicative constants. These constants play a role in the so–called *asymptotic variance* of the estimator, which is a precise descriptor of an estimator's statistical efficiency. Convergence rates and asymptotic variances have been the subject of extensive research in the statistical literature; in particular, well–known lower bounds exists regarding both the best possible (asymptotic) convergence rate of an estimator and its best possible asymptotic variance. These results set a clear frame of reference to interpret individual convergence rates, and are routinely present in the analysis of modern statistical algorithms such as noise-contrastive estimation [5, 6] or score matching [7, 8, 9].

In this work, we focus on cases where (1) the true data distribution admits a density with respect to some known base measure, and (2) the model class is parametrized by a finite-dimensional parameter. In this setting, provided that the true distribution belongs to the model class, a celebrated result in statistical estimation states that the model maximizing the average log-likelihood both achieves the best possible asymptotic convergence rate (called the *parametric rate*) and the best possible asymptotic variance, called the Cramér-Rao bound (see, e.g. [10]). While this result shows that Maximum Likelihood Estimators (MLE) are asymptotically optimal, fitting them is complicated by computational hurdles when using models with intractable normalizing constants. Such *unnormalized models* are common in the Machine Learning literature due to their high flexibility [11, 12]; their weakness however lies in the fact that expectations under these models have no unbiased approximation. For this reason, popular approximation algorithms such as unbiased gradient-based

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

stochastic optimization of the empirical log-likelihood cannot *a priori* be used, as the gradient of the normalizing constant is given by an expectation under the model distribution.

The Contrastive Divergence (CD) algorithm [13] is a popular approach that circumvents this issue by using a Markov Chain Monte Carlo (MCMC) algorithm to approximate the gradient of the log-likelihood. Unnormalized models trained with Contrastive Divergence have been shown to reach competitive performance in high-dimensional tasks such as image [14, 15, 16], text [17], and protein modeling [18, 19], or neuroscience [20]. A consistency analysis of the Contrastive Divergence algorithm is delicate, however: indeed, the *optimization error* e.g. the difference between the estimate returned by CD and the MLE, is likely to be non-negligible as compared with *statistical error* – the distance between the MLE and the true distribution – and thus cannot be discarded, as often done when analyzing estimators that minimize tractable objectives [8, 5]. Recent work [21] elegantly established asymptotic $O(n^{-1/3})$–consistency of the CD estimator for unnormalized exponential families when using only a *finite* number of MCMC steps. Key to their argument is the fact that the bias of the CD gradient estimate decreases as iterates approach the data distribution. However, as noted by the authors, their work left open the question of whether and under what conditions CD might achieve $O(n^{-1/2})$–consistency.

Contributions In this work, we answer this question by providing a non-asymptotic analysis of the CD algorithm for unnormalized exponential families. While existing convergence bounds [21] were derived for the "full batch" setting, where the CD gradient is estimated using the full dataset at each iteration, our analysis covers both the online setting (where data points are processed one at a time without replacement), and the offline setting with multiple data reuse strategies (including full batch).

In the online case (Section 3), we show, under a restricted set of assumptions compared to Jiang et al. [21], that the CD iterates can converge to the true distribution at the parametric $O(n^{-1/2})$ rate. Our analysis reveals that CD contains two sources of approximation: a bias term, and a variance term. These sources are almost independent of each other, in the sense that decreasing the bias by increasing the number of MCMC steps will not decrease the variance. The impact of these two sources of approximation transparently propagates in our resulting bounds: in particular, as the bias of the CD algorithm goes to 0, our bounds recover well-known results in online stochastic optimization [22]. Finally, we study the asymptotic variance of an estimator obtained by averaging the CD iterates, a classic acceleration technique in stochastic optimization [23]. We show that provided that the number of steps m is sufficiently large, the *asymptotic* variance of this estimator matches (up to a factor 4) the Cramér-Rao bound.

Next, we study the offline setting (Section 4), where the CD gradient is estimated by reusing (potentially random) subsets of a finite dataset. We show that a similar result to the online setup holds, up to an additional correlation term that arises from data reuse, and present several approaches to control this term. We improve over the results of [21] by showing a non-asymptotic and near-parametric rate at $O((\log n)^{1/2} n^{-1/2})$ under their conditions, and also illustrate how different rates can be obtained under a variety of conditions. Our results also show an interesting tradeoff between the effect of initialization and the statistical error as a function of batch size.

In summary, we establish the near–optimality of a variety of Contrastive Divergence algorithms for unnormalized exponential families in the so called "long-run" regime, where the number of MCMC steps is high enough to ensure that the CD gradient bias is sufficiently offset by the convexity of the negative log-likelihood.

2 Contrastive Divergence in Unnormalized Exponential Families

Unnormalized Exponential Families Exponential families (EF) [24, 25] form a well-studied class of probability distributions, given by

$$p_\psi(\mathrm{d}x) := e^{\psi^\top \phi(x) - \log Z(\psi)} c(\mathrm{d}x), \quad Z(\psi) := \int_{\mathcal{X}} e^{\psi^\top \phi(x)} c(\mathrm{d}x). \tag{1}$$

Here, $\mathcal{X} \ni x$ is the *data* or *sample space*, which we set to be a subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}^*$, although our results are readily extendable to more general measurable spaces. c is a measure on $\mathcal{X}$ called the *base* or *carrier* measure. When $\mathcal{X} \subseteq \mathbb{R}^d$, c is often set to be the corresponding Lebesgue measure. $\psi \in \Psi \subseteq \mathbb{R}^p$ is a finite-dimensional parameter called the *natural parameter*, and $\phi : \mathbb{R}^d \longmapsto \mathbb{R}^p$ is a function called the *sufficient statistics*, which, alongside with the base measure, fully describes an exponential family. Finally, $\log Z(\psi)$, the *log–normalizing* (or *cumulant*) function, is a quantity

ensuring that p_ψ integrates to 1 over $\mathcal{X}$. Crucially, we will not assume that $\log Z(\psi)$ admits a closed form expression for all ψ. The latter fact provides the practitioner with a great deal of flexibility in designing the model class: indeed, the only requirement that should be satisfied prior to performing statistical estimation is to have $Z(\psi) < +\infty$ for all ψ, something that can be readily verified and is often the case in practice. The drawback of unnormalized EFs is the fact that sampling (and thus approximating expectations under the model) cannot usually be performed in an unbiased manner. Instead, inference in unnormalized EFs is often performed using tools from the Bayesian Inference literature, such as MCMC [26]. Unnormalized EFs belong to the larger class of *unnormalized models* [27, 28, 7, 6], of the form $e^{-E_\psi(x)-\log Z(\psi)}c(\mathrm{d}x),\ Z(\psi) = \int e^{-E_\psi(x)}c(\mathrm{d}x)$, for some parametrized function $E_\psi : \mathbb{R}^d \longmapsto \mathbb{R}$ referred to as the *energy*. Unnormalized models thus take the flexibility of unnormalized EFs one step further by allowing the (negative) unnormalized log–density to be an arbitrary function E_ψ of x and ψ, instead of requiring a linear dependence on ψ as in Equation 1. We focus in this work on unnormalized EFs due to the multiple computational benefits they provide, as explained in the next section, but we believe that extending our analysis to more general unnormalized models is an interesting avenue for future work.

Statistical Estimation in Unnormalized Exponential Families using Contrastive Divergence We now review the Contrastive Divergence algorithm, an algorithm used to fit unnormalized models, and our main object of study in this work. The general setting is the following: we assume access to n i.i.d. samples $(X_1, \dots, X_n)$ drawn from some unknown distribution $p^\star$, which we assume belongs to $\mathcal{P}_\psi$, e.g. $p^\star = p_{\psi^\star}$ for some $\psi^\star \in \Psi$. Given these samples, we aim to perform *statistical estimation*, e.g. find a parameter ψ_n within Ψ that should approach $\psi^\star$ as n grows.

The starting point of the Contrastive Divergence algorithm is the unfortunate realization that Maximum Likelihood Estimation, which corresponds to minimizing the cross-entropy $\mathcal{L}(\psi) := -\mathbb{E}_{p_n} \log \mathrm{d}p_\psi/\mathrm{d}c$ between the model p_ψ and the empirical data distribution $p_n := 1/n \sum_{i=1}^n \delta_{X_i}$, cannot be performed using exact (possibly stochastic) gradient-based optimization, as the gradient $\nabla_\psi \mathcal{L}(\psi)$ of $\mathcal{L}$ with respect to the parameter ψ contains an expectation under the model distribution p_ψ. Indeed, the cross entropy and its gradient are given by

$$\begin{cases} \mathcal{L}(\psi) & = -\frac{1}{n}\sum_{i=1}^n \phi(X_i)^\top \psi + \log Z(\psi) \\ \nabla_\psi \mathcal{L}(\psi) & = -\frac{1}{n}\sum_{i=1}^n \phi(X_i) + \mathbb{E}_{p_\psi}\phi. \end{cases} \tag{2}$$

The second line follows from the well known identity $\nabla_\psi \log Z(\psi) := \mathbb{E}_{p_\psi}\phi$; we refer to [25, Proposition 3.1] for a proof. The Contrastive Divergence algorithm circumvents this issue by running approximate stochastic gradient descent (SGD) on $\mathcal{L}$, where the intractable expectation in $\nabla_\psi \log Z$ is estimated using an MCMC algorithm initialized at the empirical data distribution. In more details, given a number of epochs T, a sequence of data batches $B_{t,j}$ of size B (e.g. $B_{t,j} \in [\![1, n]\!]^B, 1 \le t \le T, 1 \le j \le N\lceil n/B \rceil$), and a family of *Markov kernels* $\{k_\psi, \psi \in \Psi\}$ each with invariant distribution p_ψ, at the jth minibatch of epoch t, $\nabla_\psi \log Z(\psi_{t,j-1})$ is approximated by $\frac{1}{B}\sum_{i \in B_{t,j}} \phi(\tilde{X}_i^m)$, where $\tilde{X}_i^m$ is produced by running the recursion $\tilde{X}_i^k \sim k_{\psi_t}(\tilde{X}_i^{k-1}, \cdot),\ \tilde{X}_i^0 = X_i$ up to $k = m$. Throughout the paper, we will refer to the conditional distribution of $\tilde{X}_i^m$ given X_i as $k_\psi^m(X_i, \cdot)$. The resulting gradient estimate arising from combining this approximation with the other (tractable) sum over the data samples present in $\nabla_\psi \mathcal{L}(\psi)$, which we refer to as the *CD gradient* and denote as h_t, is thus

$$h_{t,j} := \frac{1}{B}\sum_{i \in B_{t,j}} \phi(X_i) - \frac{1}{B}\sum_{i \in B_{t,j}} \phi(\tilde{X}_i^m) = \frac{1}{B}\sum_{i \in B_{t,j}} \left(\phi(X_i) - \phi(\tilde{X}_i^m)\right). \tag{3}$$

Key to the behavior and analysis of the CD algorithm is the strategy employed to generate minibatches $B_{t,j}$. The case where $T = 1$, $B = 1$, and $B_{1,j} = \{j\}$ will be referred to as *online* CD, while the variant where $T > 1$, and each batch $B_{t,j}$ draws B indices (with or without replacement) from $[\![1, n]\!]$ will be referred to as *offline* CD. In online CD, each data point is present in one and one batch only, while in offline CD, data points are reused across batches. From a statistical perspective, we will see that online CD can be analyzed in a remarkably simple way, while offline CD introduces additional correlations that require care to be controlled. Both settings come with their advantages and drawbacks, as we will see in the next section. The CD algorithms we study will employ decreasing step size schedules $(\eta_t)_{t \ge 0}$ of the form $\eta_t = Ct^{-\beta}$, where $C > 0$ is the initial leaning rate and $\beta \in [0, 1]$. We lay out online CD and offline CD in Algorithms 1 and 2. Note that our algorithms include a projection step on the parameter space Ψ to account for the case where Ψ is compact. In the case $\Psi = \mathbb{R}^p$, this step can be omitted. Next we depart from the setting of [21] and start by analyzing online CD.

Algorithm 1 Online CD

Input: $(X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} p_{\psi^\star}$
Parameters: Model class $\{p_\psi, \psi \in \Psi\}$, Markov kernels $\{k_\psi, \psi \in \Psi\}$, number of MCMC steps m, learning rate schedule $\eta_t := Ct^{-\beta}$, $\beta \in [0,1]$, $C > 0$, initial parameter ψ_0
for $t = 1, \dots, n$ **do**
  `//Approx. sample from` $p_{\psi_{t-1}}$
  $\tilde{X}_t^m \sim k^m_{\psi_{t-1}}(X_t, \cdot)$
  $h_t := \phi(X_t) - \phi(\tilde{X}_t^m)$
  $\psi_t \leftarrow \psi_{t-1} - \eta_t h_t$
  $\psi_t \leftarrow \text{Proj}_\Psi(\psi_t)$
end for
return ψ_n

Algorithm 2 Offline CD

Input: $(X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} p_{\psi^\star}$
Parameters: Same as Algorithm 1, plus number of epochs T, batch size B, batching schedule $B_{t,j}$, initial parameter $\psi_{0,0}$
for $t = 1, \dots, T$ **do**
  for $j = 1, \dots \lceil n/B \rceil$ **do**
    $[\tilde{X}^m_{t,i,j} \sim k^m_{\psi_{t-1}}(X_i, \cdot)$ `for i in` $B_{t,j}]$
    $h_{t,j} := \frac{1}{B} \sum_{i \in B_{t,j}} (\phi(X_i) - \phi(\tilde{X}^m_{t,i,j}))$
    $\psi_{t,j} \leftarrow \psi_{t,j-1} - \eta_t h_{t,j}$
    $\psi_{t,j} \leftarrow \text{Proj}_\Psi(\psi_{t,j})$
  end for
end for
return ψ_T

3 Non-asymptotic analysis of Online CD

3.1 Preliminaries and Assumptions

Recall that the chi-squared divergence between two probability measures p and q is defined as: $\chi^2(p,q) := \int (\frac{\mathrm{d}p}{\mathrm{d}q}(x) - 1)^2 q(\mathrm{d}x)$ if $p \ll q$, and $+\infty$ otherwise. Here, $p \ll q$ denotes that p is absolutely continuous with respect to q and $\mathrm{d}p/\mathrm{d}q$ is the Radon-Nikodym derivative [29] of p with respect to q. Let $L^2(p_\psi)$ be the space of square-integrable functions with respect to p_ψ. For a function $f \in L^2(p_\psi)$, we define

$$\alpha(f, \psi) = \frac{\left(\int \left(\int (f - \mathbb{E}_{p_\psi} f)(y)\, k_\psi(x, \mathrm{d}y)\right)^2 p_\psi(\mathrm{d}x)\right)^{1/2}}{\left(\int (f - \mathbb{E}_{p_\psi} f)(x)^2 p_\psi(\mathrm{d}x)\right)^{1/2}} \tag{4}$$

which is a measure of how quick a Markov chain with kernel k_ψ mixes, relative to the function f [30]. With these definitions in hand, we now state the assumptions required by our analysis of online CD. These assumptions form a strict subset of the assumptions considered in prior work [21], which required additional regularity and tail conditions on the Markov kernels k_ψ.

Assumption A1. $\mathcal{P}_\psi$ is a subset of a *regular* and *minimal* [25, Section 3.2] exponential family with natural parameter domain $\mathcal{D} \subseteq \mathbb{R}^p$, Ψ is a *convex and compact* subset of $\mathcal{D}$, and ψ^* lies in the interior of Ψ.

Assumption A2. There exists a constant $C_\chi > 0$ such that $\chi^2(p_{\psi^\star}, p_\psi) \leq C_\chi^2 \|\psi - \psi^\star\|^2$

Assumption A3. $\alpha := \sup\{\alpha(f, \psi),\ f \in \{\phi_i\}_{i=1}^p \cup \{\phi_i \phi_j\}_{i,j=1}^p,\ \psi \in \Psi\} < 1$, where ϕ_i is the i-th component of the function ϕ, and ϕ_i^2 is the i-th component of the function $x \longmapsto \phi(x)^2$.

A well known property of EFs [25, Proposition 3.1] is that their negative cross-entropy (against any other measure) is C^∞, convex, and strictly so if the exponential family is minimal (meaning that the set of sufficient statistic functions ϕ_i are not linearly dependent). Leaving aside the issue of intractable expectations, this convexity suggests that $\mathcal{L}$ can be efficiently minimized using stochastic approximation algorithms [31, 22]. The compactness of Ψ provided by Assumption A1 thus ensures, by the extreme value theorem [32], the existence of finite positive constants μ and L defined as:

$$\mu := \min_{\psi \in \Psi} \lambda_{\min}\left(\nabla^2_\psi \mathcal{L}(\psi)\right), \quad L := \max_{\psi \in \Psi} \lambda_{\max}\left(\nabla^2_\psi \mathcal{L}(\psi)\right), \tag{5}$$

where $\nabla^2_\psi \mathcal{L}$ is the Hessian of $\mathcal{L}$ with respect to ψ. μ (called the *strong convexity* constant) and L (a bound controlling the smoothness of the problem) play a critical role in the analysis of convex optimization algorithms [31]. While it is possible to obtain convergence rates in non-smooth or non-strongly-convex settings, our analysis follows the spirit of [21] by leveraging the strong convexity of the problem to compensate for the bias introduced by using CD gradients instead of unbiased stochastic gradients.

Assumption A2 allows link variations in distribution space to variations in parameter space, and will be instrumental to control the bias of the CD gradient. Note that since $\chi^2(p_{\psi^\star}, p_\psi) =$

$e^{\log Z(2\psi-\psi^\star)-(2\log Z(\psi)-\log Z(\psi^\star))} - 1$ provided that $2\psi - \psi^\star \in \mathcal{D}$ (see [33, Lemma 1]), we expect Assumption A2 to hold in many cases of interests. On the other hand, the possible exponential scaling of C_χ w.r.t $\log Z$ suggests that this constant may be large in some instances.

Assumption A3 is a *restricted* spectral gap condition: it guarantees that the time required by the MCMC algorithm to estimate expectations of ϕ and ϕ^2 under p_ψ will be uniformly bounded. This assumption is weaker than the (unrestricted) uniform spectral gap condition of [21], which requires that α controls the convergence rate of *all* functions in $L^2(p_\psi)$. Note that standard results in stochastic analysis [34] guarantee that $\alpha \leq 1$: thus, it only remains to ensure that α is strictly less than 1. Spectral gaps are strongly dependent on two properties of distribution: their tail behavior and their multimodality. While multimodality poses the risk of pushing the constant α close to 1, very heavy tails distributions may not verify the spectral gap condition at all.

3.2 Results

3.2.1 Parametric convergence of online CD

In this section, we show that under the assumptions stated in Section 3.1, the iterates ψ_t produced by the online CD algorithm described in Algorithm 1 will converge to the true parameter $\psi^\star$ at the parametric rate $O(n^{-1/2})$. To do so, we follow a well known paradigm in convex optimization [22] by deriving a recursion on the quantity $\delta_t \coloneqq \mathbb{E}\,\|\psi_t - \psi^\star\|^2$, which will allow, after unrolling, to obtain convergence rates for the iterates ψ_t. We aim to characterize precisely the impact of performing CD as opposed to performing online SGD on $\mathcal{L}$, which would consist of replacing the CD gradient h_t of Algorithm 1 by the unbiased (stochastic) gradient, given by:

$$g_t(\psi) \coloneqq -\phi(X_t) + \nabla_\psi \log Z(\psi) \tag{6}$$

which satisfies $\mathbb{E}\, g_t(\psi) = \nabla_\psi \mathcal{L}(\psi)$. The only stochasticity in g_t comes from the sampling of a single data point x_t from the true distribution, which is unavoidable in the online setting, and we have $\mathbb{E}\|g_t(\psi^\star)\|^2 = \mathrm{Tr}(\mathrm{Cov}_{p_{\psi^\star}}\phi) =: \sigma_\star^2$. $\sigma_\star^2$ plays a key role in the analysis of Stochastic Gradient Descent [22]. We expect that replacing g_t by h_t will introduce two sources of approximation: a bias term coming from using a finite number of MCMC steps m, and an *additional* variance term, coming from using a single sample $\tilde{x}_t^m$ to estimate $\nabla_\psi \log Z(\psi_t)$. With that in mind, we derive a recursion on δ_t in the following lemma.

Lemma 3.1. *Let $(\psi_t)_{0\leq t\leq n}$ be the iterates from Algorithm 1. Denote $\delta_t = \mathbb{E}\|\psi_t - \psi^\star\|^2$, $\sigma_\star = (\mathbb{E}_{p_{\psi^\star}}\|\phi - \mathbb{E}_{p_{\psi^\star}}\phi\|^2)^{1/2}$, and $\sigma_t = (\mathbb{E}_{p_{\psi_t}}\|\phi - \mathbb{E}_{p_{\psi_t}}\phi\|^2)^{1/2}$. Then, under A1, A2 and A3, for all $t \geq 1$,*

$$\delta_t \leq (1 - 2\eta_t\tilde{\mu}_{m,t-1} + 2\eta_t^2 L^2)\delta_{t-1} + 2\eta_t^2\tilde{\sigma}_{m,t-1}^2 + 4\alpha^{m/2}\eta_t^2\,\|\log Z\|_{3,\infty}\, C_\chi \delta_{t-1}^{1/2} \tag{7}$$

where $\|\log Z\|_{3,\infty}$ is a constant, $\tilde{\mu}_{m,t} \coloneqq \mu - \alpha^m\sigma_t C_\chi$, and $\tilde{\sigma}_{m,t} \coloneqq (\sigma_\star^2 + \sigma_t^2 + 2\sigma_t^2\alpha^{2m})^{1/2}$.

Lemma 3.1 is proved in Appendix D.2, which details the form of $\|\log Z\|_{3,\infty}$, a constant that we expect to scale roughly as dL. Loosely speaking, this recursion suggests that as the learning rate η_t goes to 0, the two terms scaling in η_t^2 will be negligible, in which case we will have: $\delta_t \leq (1 - 2\eta_t\tilde{\mu}_{m,t-1})\delta_{t-1} < \delta_{t-1}$, yielding convergence of δ_t to 0. We make these arguments formal in the next theorem. The reader familiar with the convex optimization literature will note the similarities between this recursion and the one derived in [22], which would apply as is to online SGD on $\mathcal{L}$ using g_t. The difference between the two recursions is that the roles of the strong convexity constant μ and the noise $\sigma_\star$ are now played respectively by

$$\tilde{\mu}_{m,t-1} = \mu - \alpha^m\sigma_{t-1}C_\chi \quad \text{and} \quad \tilde{\sigma}_{m,t-1}^2 = \sigma_\star^2 + \sigma_t^2 + 2\sigma_t^2\alpha^{2m}$$

These two modifications respectively characterize the impact of the bias and the additional variance introduced by the CD gradient. The last term in Equation 7, scaling in $\alpha^{m/2}\eta_t^2\sqrt{\delta_t}$, is a residual higher order mixed term coming from relating the variance of the Markov chain sample $\tilde{x}_t^m$ to σ_t^2. This term can be easily controlled as done next, and disappears as $m \to \infty$. Investigating the impact of m in the recursion, we notice that as $m \to \infty$, $\tilde{\mu}_{m,t} \to \mu$. As we will see later, this ensures that CD will converge for a sufficiently high m. On the other hand, in that same regime, $\tilde{\sigma}_{m,t}$ *does not* converge to $\sigma_\star$, but rather to $(\sigma_\star^2 + \sigma_t^2)^{1/2}$, showing the irreducible impact of the variance term.

While we precisely investigate the impact of the residual variance term in the next section, we now unify $\sigma_\star$ and σ_t by introducing

$$\sigma \coloneqq \sup_{\psi\in\Psi}(\mathbb{E}_{p_\psi}\|\phi - \mathbb{E}_{p_\psi}\phi\|^2)^{1/2}. \tag{8}$$

σ is an upper bound on the noise induced *both* by the CD gradient and by the online setup, and was used in prior work [21]. Note that by the properties of $\log Z$, σ^2 also equals $\sup_{\psi\in\Psi}\operatorname{tr}(\nabla^2_\psi\mathcal{L}(\psi))$, where $\operatorname{tr}(A)$ is the trace of $A \in \mathbb{R}^{p\times p}$, and thus finite by the extreme value theorem. The following theorem is obtained by invoking standard unrolling arguments in the convex optimization literature. In the next result, we use the function $\varphi_\gamma(t)$, defined as $\varphi_\gamma(t) = \frac{t^\gamma - 1}{\gamma}$ if $\gamma \neq 0$, and $\log t$ if $\gamma = 0$.

Theorem 3.2. *Fix $n \geq 1$. Let $(\psi_t)_{0\leq t\leq n}$ be the iterates produced by Algorithm 1, and define $\delta_t \coloneqq \mathbb{E}\|\psi_t - \psi^\star\|^2$. Moreover, assume that $m > \frac{\log(\sigma C_\chi/\mu)}{\log|\alpha|}$, i.e. $\tilde{\mu}_m \coloneqq \mu - \alpha^m\sigma C_\chi > 0$. Then under Assumptions A1, A2 and A3, for $\eta_t = Ct^{-\beta}$ with $C > 0$, we have:*

$$\delta_n \leq \begin{cases} 2\exp\left(4\tilde{L}C^2\varphi_{1-2\beta}(n)\right)\exp\left(-\frac{\tilde{\mu}_m C}{4}n^{1-\beta}\right)\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + \frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m n^\beta}, & \text{if } 0 \leq \beta < 1 \\ \frac{\exp\left(2\tilde{L}^2C^2\right)}{n^{\tilde{\mu}C}}\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + 2\tilde{\sigma}_m^2 C^2\frac{\varphi_{\tilde{\mu}_m C/2-1}(n)}{n^{\tilde{\mu}_m C/2}}, & \text{if } \beta = 1\ , \end{cases}$$

where $\tilde{\sigma}_m = \sigma^2(2 + 2\alpha^{2m}) + \alpha^{m/2}\|\log Z\|^2_{3,\infty}C_\chi^2$ and $\tilde{L} = (L^2 + \alpha^{m/2})^{1/2}$. Consequently, if $\eta_n = \frac{C}{n}$ with an initial learning rate $C > 2\tilde{\mu}_m^{-1}$, we have $\sqrt{\delta_n} \leq 2\tilde{\sigma}_m C\sqrt{\frac{\tilde{\mu}_m C}{\tilde{\mu}_m C - 2}}\frac{1}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$.

Theorem 3.2 is proved in Appendix D.3. It shows that the iterates produced by online CD will converge to the true parameter $\psi^\star$ at the rate $O(n^{-1/2})$ provided that the number of steps m is sufficiently large, improving over the asymptotic $O(n^{-1/3})$ rate of [21], while imposing slightly weaker conditions on the number of steps m (see [21, Theorem 2.1]). This proves that online CD can be asymptotically competitive with other methods for training unnormalized models, such as Noise Contrastive Estimation [6], or Score Matching [7]. However, the asymptotic variance of ψ_t (e.g. the multiplicative factor in front of the $O(n^{-1/2})$ term) is likely to be suboptimal, e.g. much larger than the Crámer-Rao bound, given by the trace of the inverse of the Fisher information matrix [25]. Given the statistical optimality of MLE, and the fact that CD in an approximate MLE method, this motivates the further goal or obtaining a CD estimator with near-optimal statistical properties. In the next section, we achieve this goal by showing that averaging the iterates ψ_t will produce a near statistically-optimal estimator, in a sense that we will make precise.

3.2.2 Towards statistical optimality with averaging

Polyak-Ruppert averaging [23] is a simple yet surprisingly effective way to construct an asymptotically optimal estimator $\bar{\psi}_n \coloneqq \frac{1}{n}\sum_{t=1}^n \psi_i$ from a sequence of iterates $(\psi_t)_{0\leq t\leq n}$ obtained by running a standard online SGD algorithm [22]. As shown in [22], when the objective is the cross-entropy of a model, and assuming the unbiased stochastic gradients are available, averaging yields an estimator $\overline{\psi}$ with the asymptotic variance $\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})/n$, where $\mathcal{I}(\psi) \coloneqq \operatorname{Cov}_{p_{\psi^\star}}\phi$ is the Fisher information matrix of the data distribution $p_{\psi^\star}$. $\mathcal{I}(\psi^\star)^{-1}$ being the Cramér-Rao *lower bound* on asymptotic variances of statistical estimators [10], this estimator $\bar{\psi}_n$ is asymptotically optimal. The following theorem shows conditions under which averaging CD iterates can give rise to a near-optimal estimator.

Theorem 3.3 (Contrastive Divergence with Polyak-Ruppert averaging)**.** *Let $(\psi_t)_{t\geq 0}$ the sequence of iterates obtained by running the CD algorithm with a learning rate $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$. Define $\bar{\psi}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \psi_i$. Then, under the same assumptions as Theorem 3.2, and assuming additionally that $m \coloneqq m(n) > \frac{(1-\beta)\log n}{2|\log\alpha|}$, we have, for all $n \geq 1$,*

$$\left(\mathbb{E}\left\|\overline{\psi}_n - \psi^\star\right\|^2\right)^{1/2} \leq 2\sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + o(n^{-1/2})$$

Consequently, we have that $\limsup_{n\to\infty} n\mathbb{E}(\left\|\overline{\psi}_n - \psi^\star\right\|^2) \leq 4\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})$.

Theorem 3.3, alongside with a statement which includes the asymptotic order of the residual term, is proved in Appendix D.4. It shows that at the cost of an increase in *computational* complexity of the entire algorithm from $O(n)$ to $O(n\log n)$, $\bar{\psi}_n$ will be a near-optimal statistical estimator of $\psi^\star$.

While this increase in complexity emerges from the bias of CD, the additional variance of CD results in an asymptotic variance inflated by a factor of 4 compared to the Cramér-Rao bound.

Theorem 3.3 concludes our analysis of online CD. Despite their asymptotic near-optimality, the bounds provided for online CD and its averaged version have weaknesses: the online CD iterates are not robust to choices of C. On the other hand, as shown in Appendix D.4, the bound of the averaged iterates contain higher-order terms that could be large in intermediate sample regimes. Next, we show that offline CD, which processes data points multiple times, can alleviate these issues.

4 Non-asymptotic analysis of offline CD

In practice, CD gradient approximation schemes are commonly used within an offline stochastic gradient descent (SGD) algorithm, where one is given the full size-n dataset upfront and each update uses some stochastic subset of the data. We study CD under offline SGD with replacement (SGDw), i.e. Algorithm 2 with batches $B_{t,j}$ being i.i.d. uniform draws of size-B subsets of $[n]$, and include SGD without replacement in Appendix B.2. To do so, we follow the setting of prior work on offline CD [21], which established its asymptotic $O(n^{-\frac{1}{3}})$ consistency. We show that by slightly strengthening a moment assumption used in [21], the offline CD iterates converge to the true parameter at a near-parametric $O((\log n)^{\frac{1}{2}} n^{-\frac{1}{2}})$ rate. Our proof proceeds by controlling a "tail probability" term specific to the offline setting which characterizes the strength of the correlations between the offline CD iterates and the training data. While, as we show, the assumptions of [21] provide a tail control sufficient to obtain a near-parametric rate, other strategies are possible to obtain convergence guarantees. In particular, we show that non-asymptotic convergence can be obtained by either (1) relaxing assumptions on the Markov kernel required by prior work, or (2) making a specific mixing assumption the Markov chain.

4.1 Background: Asymptotic consistency of offline CD in subexponential settings

Prior work [21] has established *asymptotic* $O(n^{-\frac{1}{3}})$ consistency of the (averaged) offline CD iterates in the full-batch case. We summarize their results and assumptions below.

Assumption A4. There exists $\nu \geq 2$ s.t. for all $m \in \mathbb{N}$, there is $\kappa_{\nu;m} < \infty$ s.t.

$$\sup_{x\in\mathcal{X}} \sup_{\psi\in\Psi} \left(\mathbb{E}\left\|\phi(K_\psi^m(x)) - \mathbb{E}[\phi(K_\psi^m(x))]\right\|^\nu\right)^{1/\nu} \leq \kappa_{\nu;m} .$$

Assumption A5. There exists some $C_m > 0$ such that, for all $\psi_1, \psi_2 \in \Psi$, $\sup_{x\in\mathcal{X}} \|\mathbb{E}[\phi(K_{\psi_1}^m(x))] - \mathbb{E}[\phi(K_{\psi_2}^m(x))]\| \leq C_m\|\psi_1 - \psi_2\|$.

Assumption A6. There exist some $\sigma_m, \zeta_m > 0$ such that, for any $z \in \mathbb{R}^p$ with $\|z\| \leq \zeta_m$, $\mathbb{E}[e^{z^\top(\phi(K_{\psi^*}^m(X_1)) - \mathbb{E}[\phi(K_{\psi^*}^m(X_1))])}] \leq e^{\sigma_m^2\|z\|^2/2}$.

Theorem 4.1 (Theorem 2.1 of [21]). *Assume assumptions A1,A2, A3, A4 (for $\nu = 2$), A5 and A6. Let $\psi_{t,1}$ be the t-th iterate of offline CD with full-batch gradient descent and constant stepsize $\eta_t = C$, i.e the iterates produced by Algorithm 2 using $B_{t,1} = [\![1, n]\!]$. Then for any learning rate C and number of Markov kernel steps m satisfying $\mu - \alpha^m\sigma C_\chi - \frac{C}{2}(L + \alpha^m\sigma C_\chi)^2 > 0$, we have, for some $A_m > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\limsup_{T\to\infty} \left\|\frac{1}{T}\sum_{t=1}^{T} \psi_{t,1}^{\mathrm{SGDw}} - \psi^\star\right\| > A_m n^{-\frac{1}{3}}\right) = 0$$

This result shows convergence of the *averaged* full-batch CD iterates to the true parameter in the large n and T limit. As discussed, this result is asymptotic both in n and T: the probability of the error exceeding $A_m n^{-\frac{1}{3}}$ goes to 0 as $n \to \infty$ and $T \to \infty$, but at an unknown rate. Moreover, the $O(n^{-\frac{1}{3}})$ does not match the optimal $O(n^{-\frac{1}{2}})$ rate.

4.2 Sharpening offline CD bounds in subexponential settings

4.2.1 Non-asymptotic $\tilde{O}(n^{-1/2})$-consistency

As a first result, we show that under the assumptions of [21] (except for a slightly stronger $\nu > 2$ moment assumption in A4), $\psi_{T,N}^{\mathrm{SGDw}}$ in fact achieves a near-parametric rate. The most general version of our result holds for any learning rate schedule of the form $Ct^{-\beta}, \beta \in [0, 1]$, and for offline SGD

with arbitrary batch sizes B, with data drawn either with or without replacement across batches. For simplicity, we first present our result assuming full batch ($B = n, N = 1, \psi_{t,j}^{\mathrm{SGDw}} = \psi_{t,1}^{\mathrm{SGDw}}$ for $t \geq 1$) SGD with constant step sizes $\eta_t = C$, which is the setting of [21]. Analogue bounds holding for the other mentioned batching and step sizes schedules can be found in Appendix B.

Theorem 4.2. *Assume the setup of Theorem 4.1, except that Assumption A4 holds for some $\nu > 2$, and that $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 4CL^2$. Let $\delta_{t,j}^{\mathrm{SGDw}} := \mathbb{E}\|\psi_{t,j}^{\mathrm{SGDw}} - \psi^*\|^2$. Then, we have:*

$$\sqrt{\delta_{T,1}^{\mathrm{SGDw}}} \leq E_1^{T,1}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} + C'(p,\nu,m,\Psi)\Big(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\Big)\Big(\frac{e^{\frac{\tilde{\mu}_m C}{2}}}{\tilde{\mu}_m C} + \frac{E_2^{T,1}}{L^2C^2}\Big) \tag{9}$$

where $E_1^{T,1}$, $E_2^{T,1}$ are functions decreasing exponentially in T, and $C'(p,\nu,m,\Psi)$ is a constant in n, T. Consequently,

$$\lim_{T\to\infty}\sqrt{\delta_{T,1}^{\mathrm{SGDw}}} \leq \frac{e^{\frac{\tilde{\mu}_m C}{2}}}{\tilde{\mu}_m C} C'(p,\nu,m,\Psi,\beta)\Big(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\Big) .$$

The precise values of all the constants can be found in Theorem B.1 (for $E_1^{T,1}$, $E_2^{T,1}$) and Lemma B.3 (for $C'(p,\nu,m,\Psi,\beta)$), including their expressions for $N > 1$ and $\beta \in [0,1]$. We comment on the main differences between our result and the one of [21]. First our bound holds for *any* epoch T and number of samples n. Second, fixing n but taking $T \to \infty$, the final bound matches the parametric $O(\sqrt{n})$ up to a $\sqrt{\log(n)}$ factor, a significant improvement over the $O(n^{-\frac{1}{3}})$ rate of [21]. Finally, we control an L_2 error, which is a stronger control than a high probability bound by Markov's inequality; we hypothesize this is the reason why a slightly stronger moment assumption is required for our setup, compared to the one used for the high probability bound in [21].

Inspecting Equation 9, we notice the presence of two *transient* terms, and a *stationary term*, reminiscent of the structure of upper bound of Theorem 3.2. The transient terms (i.e. the ones containing $E_1^{T,1}$ and $E_2^{T,1}$) vanish exponentially fast in the total number of CD updates T. However, unlike in online CD where the number of updates and the number of samples are tied (e.g. $T = n$), these two values are now *decoupled*, and these terms can be made arbitrarily small by increasing the number of gradient steps T without having to collect more samples n. The stationary term, which is the only one remaining in the limit of $T \to \infty$, decreases with n at a rate that is independent of hyperparameters like the step size C or the learning rate schedule β (see Lemma B.3). In that sense, offline CD compares favorably to online CD, whose rate is sensitive to β and C, and averaged online CD, whose bound contains higher-order (in n) terms which can be large in the moderate n regime. On the other hand, the stationary term in offline CD is asymptotically suboptimal: its rate is larger (while only up to a log factor) than the best-case $O(\sqrt{n})$ one achieved by online CD algorithms, and the leading constant does not match the optimal one.

4.2.2 Proof of Theorem 4.2

The high-level proof of Theorem 4.2 follows a similar strategy as the online one: first, derive a recursion for the quantity $\delta_{t,1}^{\mathrm{SGDw}} := \mathbb{E}\|\psi_{t,1}^{\mathrm{SGDw}} - \psi^*\|^2$, then unroll it explicitly to obtain a final bound on $\delta_{T,1}^{\mathrm{SGDw}}$. The main difference to online CD is the presence of an additional offline-specific correlation between the iterates and the data. We thus break down the proof into three steps: (1) deriving a controllable, uniform-in-time upper bound of the data-iterate correlations, (2) deriving and unrolling a recursion on $\delta_{t,1}^{\mathrm{SGDw}}$ containing this new term, and (3) controlling that term to obtain a final bound on $\delta_{T,1}^{\mathrm{SGDw}}$.

Step 1:characterizing the data-iterate correlations in offline CD In offline CD, at each epoch $t \geq 1$, the iterate $\psi_{t-1,1}^{\mathrm{SGDw}}$ and the data samples X_i are correlated: this is because these samples may have been used in previous epochs $t' < t-1$ to obtain the $\psi_{t',1}^{\mathrm{SGDw}}$, which themselves influenced $\psi_{t-1,1}$. With such correlations, we now have $\mathbb{P}(X_i|\psi_{t-1,1}^{\mathrm{SGDw}}) \neq \mathbb{P}(X_i)$, preventing us from obtaining an unrollable recursion on $\delta_{t,1}^{\mathrm{SGDw}}$ by first marginalizing X_i out to obtain an upper bound of $\mathbb{E}\left[\left\|\psi_{t,1}^{\mathrm{SGDw}} - \psi^*\right\|^2 | \psi_{t-1,1}^{\mathrm{SGDw}}\right]$ that only depends on $\left\|\psi_{t-1,1}^{\mathrm{SGDw}} - \psi^\star\right\|$, and then marginalizing over $\psi_{t-1,1}^{\mathrm{SGDw}}$ to obtain a recursion as in Lemma 3.1. As this problem would not have occurred had we used "fresh samples" (e.g. i.i.d copies of X_i not present in the training data) to perform our update, the

core of the proof lies in controlling the following quantity:

$$\Delta(\psi_{t,1}^{\mathrm{SGDw}}) := \Big\| \frac{1}{n} \sum_{i \le n} \big(\mathbb{E}\big[\phi\big(K^m_{i;\psi_{t,1}^{\mathrm{SGDw}}}(X_i)\big) \big| \psi_{t,1}^{\mathrm{SGDw}}, X_i\big] - \mathbb{E}\big[\phi\big(K^m_{i;\psi_{t,1}^{\mathrm{SGDw}}}(X_1')\big) \,\big|\, \psi_{t,1}^{\mathrm{SGDw}}\big]\big) \Big\|$$

where X_1' is an i.i.d. copy of X_1. $\Delta(\psi_{t,1}^{\mathrm{SGDw}})$ is the expected (over the data and iterates) error between a quantity that allows to obtain a recursion (the rightmost term) and the one actually used by offline CD (the leftmost term). To control it, we upper-bound it using a tail decomposition:

$$\mathbb{E}[\Delta(\psi_{t,1}^{\mathrm{SGDw}})^2] \le \epsilon^2 + (\sup_t \mathbb{E}[\Delta(\psi_{t,1}^{\mathrm{SGDw}})^\nu])^{2/\nu} \sup_t \mathbb{P}(\Delta(\psi_{t,1}^{\mathrm{SGDw}}) > \epsilon)^{\frac{\nu-2}{\nu}} := \varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon)^2 \tag{10}$$

We invoke an additional assumption to ensure that $(\mathbb{E}[\Delta(\psi_{t,1}^{\mathrm{SGDw}})^\nu])^{2/\nu}$ is finite; in the results of [21], this is automatically implied by assumptions A4 and A6. For simplicity we assume the same bounding constant $\kappa_{\nu;m}$.

Assumption A7. There exists $\nu \ge 2$ s.t. for all $m \in \mathbb{N}$, $\kappa_{\nu;m}$ from A4 moreover verifies

$$\sup_{\psi \in \Psi} \left(\mathbb{E}\left\| \phi(K_\psi^m(X_1)) - \mathbb{E}[\phi(K_\psi^m(X_1))] \right\|^\nu \right)^{1/\nu} \le \kappa_{\nu;m} .$$

Note the similarity of this assumption with assumption A4: the only difference is that X_1 is now a random training point instead of an deterministic (arbitrary) one. Ensuring assumption A7 in addition to assumption A4 thus requires controlling a ν-th order moment, instead of all moments as implied by assumption A6.

Step 2: Deriving and unrolling the recursion on $\delta_{t,1}^{\mathrm{SGDw}}$ The right-hand side of Equation (10) does not depend on t, allowing for the derivation of an "unrollable" recursion on $\delta_{t,1}^{\mathrm{SGDw}}$ and its subsequent unrolling, which is performed in the following theorem. For simplicity, we again assume $\beta = 0$ and $N = 1$ and defer the general case to Theorem B.1 in appendix.

Theorem 4.3 (Convergence up to a tail control)**.** *Assume A1, A2, A3, A4 and A7. Let $\eta_t = C$ for some $C > 0$, and assume that $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 4CL^2$. Then for any $\epsilon > 0$,*

$$\sqrt{\delta_{T,1}^{\mathrm{SGDw}}} \le E_1^{T,1} \sqrt{\delta_{0,0}^{\mathrm{SGDw}}} + C \left(\varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu,m}}{\sqrt{n}} \right) \left(\frac{e^{\frac{\tilde{\mu}_m C}{2}}}{\tilde{\mu}_m C} + \frac{E_2^{T,1}}{L^2 C^2} \right)$$

where $\varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon)$ is defined in Equation (10)*.*

Note that in the general, non-full batch $B \le n$ case, $\frac{5\sigma+5\kappa_{\nu,m}}{\sqrt{n}}$ is replaced by $\frac{5\sigma+5\kappa_{\nu,m}}{\sqrt{B}}$ (see Theorem B.1). Under our bounds, obtaining consistency thus requires setting $B \equiv B(n) \underset{n\to\infty}{\to} +\infty$.

Step 3: Controlling the tail probability term Theorem 4.3 is just one step away from the final bound of Theorem 4.2: it remains to control the tail term $\varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon)$. Under the assumptions of [21], minimizing $\varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon)$ over ϵ yields the following result:

Lemma 4.4. *Assume the setup of Theorem 4.2. Let $n \in \mathbb{N}$ be sufficiently large s.t. $\frac{\log n}{n} < \frac{\sigma_m^2 \zeta_m^2}{p+\nu-2}$. Denote r_Ψ as the radius of the smallest sphere in $\mathbb{R}^p$ that contains Ψ, which is finite under A1. Then*

$$\inf_{\epsilon>0} \varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon) \le$$

$$\left(\frac{3\sigma_m \sqrt{p((\nu-2)p+2\nu)}}{\sqrt{\nu-2}} + \kappa_{\nu;m} 2^{\frac{\nu-2}{2\nu}} (r_\Psi)^{\frac{(\nu-2)p}{2\nu}} \left(1 + \frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2}((\nu-2)p+2\nu)^{1/2}} \right)^{\frac{(\nu-2)p}{2\nu}} \right) \frac{\sqrt{\log n}}{\sqrt{n}}.$$

To obtain this result, we control the moment term $(\sup_t \mathbb{E}[\Delta(\psi_{t,1}^{\mathrm{SGDw}})^\nu])$ using A7, and we control the tail probability term $\sup_t \mathbb{P}(\Delta(\psi_{t,1}^{\mathrm{SGDw}}) > \epsilon)$ as in [21, Lemma 3.1] using an union bound, a covering argument and A6. Theorem 4.2 then follows by plugging Lemma 4.4 into Theorem 4.3.

4.3 Consistency of offline CD: beyond subexponential tails.

As discussed above, the general unrolling result of Theorem 4.3 holds without the subexponentiality assumption A6; this assumption was only used in Lemma 4.4 to control $\varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon)$. We now discuss two alternative ways to control this quantity without requiring subexponential tails. The first generalizes the idea of Jiang et al. [21], while the second exploits mixing of the Markov chain $K_\psi^m(x)$ as $m \to \infty$. As before we only state partial results (full batch, $\beta = 0$) and defer the full explicit bounds to Appendix B.3.

Control via Markov Inequality Our first alternative uses Markov Inequality to yield the following.

Theorem 4.5. *Assume the setup of Theorem 4.3 and additionally that A5 holds. Then*

$$\inf_{\epsilon>0} \varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon) \ \leq \tilde{C}(p,\nu,m,\Psi)\, n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}} \ ,$$
$$\lim_{T\to\infty} \sqrt{\delta^{\mathrm{SGDw}}_{T,1}} \ \leq \tilde{C}'(p,\nu,m,\Psi)\big(n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}} + \frac{1}{\sqrt{n}}\big),$$

where $\tilde{C}$ and $\tilde{C}'$ are functions whose explicit expressions are given in Lemma B.4 in the appendix.

In the case $p = 1$ and $\nu = 3$, the sub-optimal error from Theorem 4.5 reads $O(n^{-3/20})$. Theorems 4.2 and 4.5 reveal that, depending on the tail condition imposed on the noise introduced by the Markov kernel, the convergence rate of offline CD varies: A subexponential tail, as assumed in prior work, in fact leads to near-parametric rate. Meanwhile, consistency can be obtained without assuming subexponentiality, albeit at a sub-optimal rate.

Control via Markov chain mixing. Alternatively, notice that $\mathbb{E}[\Delta(\psi^{\mathrm{SGDw}}_{t,1})^2]$ involves an average of $\mathbb{E}\big[\phi\big(K^m_{\psi^{\mathrm{SGDw}}_{t,1}}(X_i)^2\big)\big|X_i, \psi^{\mathrm{SGDw}}_{t,1}\big] - \mathbb{E}\big[\phi\big(K^m_{\psi^{\mathrm{SGDw}}_{t,1}}(X'_1)\big)\big|\psi^{\mathrm{SGDw}}_{t,1}\big]$. When $m \to \infty$, the effect of initialization vanishes, and one may expect the difference to converge to zero. We defer to Lemma B.5 in the appendix to show that, under a ϕ-discrepancy mixing condition ([35]) with a mixing coefficient $\tilde{\alpha} \in [0, 1)$,

$$\inf_{\epsilon>0} \varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon) \ = O(\kappa_{\nu;m}\tilde{\alpha}^{\frac{(\nu-2)m}{3\nu-2}}) \quad \text{and} \quad \lim_{T\to\infty} \sqrt{\delta^{\mathrm{SGDw}}_{T,1}} \ = O\Big(\kappa_{\nu;m}\tilde{\alpha}^{\frac{(\nu-2)m}{3\nu-2}} + \frac{\sigma+\kappa_{\nu;m}}{\sqrt{n}}\Big) \ .$$

As $m \to \infty$, this recovers the parametric rate $O(n^{-1/2})$. This alternative convergence guarantee comes at the cost of requiring m, the number of Markov chain steps, to grow with the sample size n.

Remark (Examples)**.** In our main results (Theorems 3.2, 3.3 and 4.3) and the tail condition for offline SGD (Theorem 4.2), we employed a weaker set of assumptions than those in [21] (except for the mild $\nu > 2$ moment assumption in (A4)). Consequently, our results apply to all three examples studied in [21]: A bivariate Gaussian model with unknown mean and random-scan Gibbs sampler, a fully visible Boltzmann machine with random-scan Gibbs sampler, and an exponential-family random graph model with a Metropolis-Hastings sampler.

5 Related Work

Central to this paper is the prior work of Jiang et al. [21], which provided a rigorous theoretical foundation to analyze the convergence of full-batch CD, and which we refine. The study of optimization with biased gradient descent has attracted a lot of attention in recent years [36, 37, 38, 39]. These works, while closely connected to ours, analyze algorithms with different implementation choices than the CD algorithm: i.i.d. noise setup [36], or setup where a persistent Markov chain is maintained through the iterations [36, 37, 38, 39]. The latter is akin to a variant of the CD algorithm, called the persistent CD [40]. In contrast, our analysis focus on the CD algorithm that restarts a batch of Markov chains from the data distribution at every iteration. Finally, there is a rich body of work on convergence guarantees for offline multi-pass SGD [41, 42, 43, 44, 45, 46]. A notable difference of our analysis is that we are primarily concerned with statistical errors associated with convergence to the true parameter ψ^* in number of samples n, and not the commonly studied convergence rate in number of epochs T. Consequently, most of our work for the offline setup goes into handling the correlations that accumulate by reusing data across epochs.

6 Discussion

In this work, we provide a non-asymptotic analysis of the Contrastive Divergence algorithms, showing, in the online setting, their potential to converge at the parametric rate and to have near-optimal asymptotic variance, and proving a near-parametric rates in the offline setting, significantly extending prior results. Our results apply to unnormalized exponential families: despite their flexibility, these models only cover log-densities with linear relationships on the model parameters. We believe that extending our results to more general forms of unnormalized models is an important direction for future work.

Acknowledgments and Disclosure of Funding

All authors acknowledge support from the Gatsby Charitable Foundation.

References

[1] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[4] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[5] Holden Lee, Chirag Pabbaraju, Anish Prasad Sevekari, and Andrej Risteski. Pitfalls of gaussians as a noise distribution in nce. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010.

[7] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

[8] Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2022.

[9] Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

[11] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

[12] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002.

[14] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[15] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 2019.

[16] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy- based models. *arXiv preprint arXiv:2021.01316*, 2021.

[17] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551, 2022.

[18] Deqian Kong, Bo Pang, Tian Han, and Ying Nian Wu. Molecule design by latent space energy-based modeling and gradual distribution shifting. In *Uncertainty in Artificial Intelligence*, pages 1109–1120. PMLR, 2023.

[19] Csilla Varnai, Nikolas S Burkoff, and David L Wild. Efficient parameter estimation of generalizable coarse-grained protein force fields using contrastive divergence: a maximum likelihood approach. *Journal of chemical theory and computation*, 9(12):5718–5733, 2013.

[20] Pierre Glaser, Michael Arbel, Arnaud Doucet, and Arthur Gretton. Maximum likelihood learning of energy-based models for simulation-based inference. 2022.

[21] Bai Jiang, Tung-Yu Wu, Yifan Jin, and Wing H Wong. Convergence of contrastive divergence algorithm in exponential family. *The Annals of Statistics*, 46(6A):3067–3098, 2018.

[22] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[23] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[24] Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

[25] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[26] Charles J Geyer. Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo*, 20116022:45, 2011.

[27] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[28] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

[29] Joseph L Doob. *Measure theory*, volume 143. Springer Science & Business Media, 2012.

[30] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[31] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[32] Gary Harris and Clyde Martin. Shorter notes: The roots of a polynomial vary continuously as a function of the coefficients. *Proceedings of the American Mathematical Society*, pages 390–392, 1987.

[33] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2013.

[34] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

[35] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. Function-specific mixing times and concentration away from equilibrium. 2020.

[36] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.

[37] Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.

[38] Thinh T Doan. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2022.

[39] Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

[40] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, 2008.

[41] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

[42] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[43] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[44] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

[45] Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.

[46] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Risk bounds of multi-pass sgd for least squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 35:12909–12920, 2022.

[47] Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.

[48] Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.

[49] Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

[50] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[51] S. W. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *Ann. Math. Statist.*, pages 1719–1723, 1968.

Supplementary Material for "Near-Optimality of Contrastive Divergence Algorithms"

The supplementary material provides the proofs of the main results of the paper:

Section B states full explicit bounds for the offline CD algorithm.

Section C collects a list of useful tools for our proofs. These include the properties of φ_γ introduced before Theorem 3.2 in the main text, as well as several contraction and integrability results.

Section D provides the proofs for the online CD algorithm.

Sections E, F and G contain the proofs about the offline CD algorithm and the tail control.

A Notations

Throughout the proofs, we will denote by P_ψ^m the following operator from $L^2(p_\psi)$ to itself:

$$P_\psi^m f(x) \coloneqq \int k^m(x,x')f(x')p_\psi(x')\mathrm{d}x'. \tag{11}$$

Here, $k^m(x,x')$ is the m-iterated version of some Markov transition kernel k_ψ, e.g.:

$$k_\psi^m(x,x') \coloneqq \int k_\psi(x,x_1)\dots k_\psi(x_{m-2},x_{m-1})\dots k_\psi(x_{m-1},x')\mathrm{d}x_1\dots\mathrm{d}x_{m-1}. \tag{12}$$

$\mathrm{Proj}_\Psi : \mathbb{R}^p \longmapsto \Psi$ denotes the projection operator onto the convex set Ψ, e.g.

$$\mathrm{Proj}_\Psi(\psi) \coloneqq \underset{\psi'\in\Psi}{\mathrm{argmin}} \left\|\psi-\psi'\right\|.$$

We also frequently use the following function, used in standard convex optimization results [22].

$$\varphi_\gamma(t) = \begin{cases} \frac{t^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0 \\ \log t & \text{if } \gamma = 0 \end{cases}$$

which is defined on $\mathbb{R}_+ \setminus \{0\}$.

B Additional results for offline SGD

In this section, we provide the full statements on error bounds for SGD with replacement (SGDw), SGD with reshuffling (SGDo) and tail moment bounds, which complement the results in Section 4. Proofs are deferred to Appendix F, which make use of L_2 approximation by auxiliary gradient updates derived in Appendix E.

Notations Denote the SGDw iterates by $(\psi_{t,j}^{\mathrm{SGDw}})_{t\in\mathbb{N},j\leq N}$ and let X_1' be an i.i.d. copy of X_1. Throughout the remaining of the appendix, we define given $\epsilon > 0$ and $n \in \mathbb{N}$

$$\begin{aligned}\vartheta_{n,m,T}^{\mathrm{SGDw}}(\epsilon) &\coloneqq \sup_{\substack{t\in[T]\\ j\in[N]}} \mathbb{P}\left(\frac{\left\|\sum_{i=1}^n \left(\mathbb{E}\left[\phi\left(K^m_{\psi_{t-1,j}^{\mathrm{SGDw}}}(X_i)\right)\middle| X_i, \psi_{t-1,j}^{\mathrm{SGDw}}\right] - \mathbb{E}\left[\phi\left(K^m_{\psi_{t-1,j}^{\mathrm{SGDw}}}(X_1')\right)\middle|\psi_{t-1,j}^{\mathrm{SGDw}}\right]\right)\right\|}{n} > \epsilon\right). \\ &= \sup_{t,j} \mathbb{P}(\Delta(\psi_{t,j}^{\mathrm{SGDw}}) > \epsilon)\end{aligned}$$

and $\vartheta_{n,m,T}^{\mathrm{SGDo}}(\epsilon)$ analogously. Using these notations, we can redefine the quantity $\varepsilon_{n,m,T;\nu}^{\mathrm{SGDw}}(\epsilon)$ in the main as $\varepsilon_{n,m,T;\nu}^{\mathrm{SGDw}}(\epsilon) \coloneqq \sqrt{\epsilon^2 + \kappa_{\nu;m}^2 \left(\vartheta_{n,m,T}^{\mathrm{SGDw}}(\epsilon)\right)^{\frac{\nu-2}{\nu}}}$.

B.1 An explicit finite-sample bound for SGDw

In the result below, we write $\delta_{t,j}^{\mathrm{SGDw}} := \mathbb{E}\left\|\psi_{t,j}^{\mathrm{SGDw}} - \psi^*\right\|^2$ and, for a fixed $\epsilon > 0$, the quantity

$$\sigma_{n,T}^{\mathrm{SGDw}} \;=\varepsilon_{n,m,T;\nu}^{\mathrm{SGDw}}(\epsilon) + \frac{5\sigma + 5\kappa_m}{\sqrt{B}} \;=\; \sqrt{\epsilon^2 + \kappa_m^2\left(\vartheta_{n,m,T}^{\mathrm{SGDw}}(\epsilon)\right)^{\frac{\nu-2}{\nu}}} + \frac{5\sigma + 5\kappa_m}{\sqrt{B}} \;.$$

Theorem B.1. *Assume A1 (where Ψ may be non-compact), A2, A3, A4 and A7. Let $\eta_t = Ct^{-\beta}$ for some $\beta \in [0,1]$ and $C > 0$, and assume that $m > \frac{\log(\sigma C_\chi/\mu)}{\log|\alpha|}$ s.t. $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 0$ as in Theorem 3.2. Then for any $\epsilon > 0$, $\sqrt{\delta_{T,N}^{\mathrm{SGDw}}}$ is upper bounded by*

$$\begin{cases} E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} \;+\; C\sigma_{n,T}^{\mathrm{SGDw}}\Big(\frac{4e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}}}{\tilde{\mu}_m C} + 2N(1+\tilde{\mu}_m C)^{N-1}\varphi_{\frac{1}{2}-L^2C^2N}(T+1)\,E_2^{T,N}\Big) & \text{for } \beta = \frac{1}{2}\,, \\ E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} + C\sigma_{n,T}^{\mathrm{SGDw}}\Big(\frac{4}{\tilde{\mu}_m C} + \frac{3N\big(1+\frac{L^2C^2}{2}\big)^{N-1}e^{2L^2C^2N}\,\log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}}\Big) & \text{for } \beta = 1\,, \\ E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} \;+\; C\sigma_{n,T}^{\mathrm{SGDw}}\Big(\frac{2^{2\beta+1}}{\tilde{\mu}_m C}e^{\frac{\tilde{\mu}_m C}{2(1-\beta)}}\frac{N}{(T+1)^\beta} + \frac{3^\beta(1+\tilde{\mu}_m C)^{N-1}(T+2)^\beta}{L^2C^2}\,E_2^{T,N}\Big) & \text{otherwise}\,, \end{cases}$$

where $E_1^{T,N}$ and $E_2^{T,N}$ are two decreasing functions in T defined by

$$E_1^{T,N} \;:=\; \exp\Big(1 - N\tilde{\mu}_m C\varphi_{1-\beta}(T+1) + \frac{NL^2C^2}{2}\varphi_{1-2\beta}(T+1)\Big)\,,$$
$$E_2^{T,N} \;:=\; \exp\Big(-\frac{N\tilde{\mu}_m C}{2}\varphi_{1-\beta}(T+1) + 2NL^2C^2\varphi_{1-2\beta}(T+1)\Big)\,.$$

We emphasize that the full result above holds for any $\beta \in [0,1]$, which in particular includes the constant step size $\beta = 0$ regime considered by [21]. When $\beta = 0$, for $E_1^{T,N}$ and $E_2^{T,N}$ to decay to zero as $T \to \infty$, we additionally need the condition

$$\tilde{\mu}_m \;=\; \mu - \alpha^m\sigma C_\chi \;>\; 4CL^2\,.$$

This is almost identical to the condition used in [21, Equation 2.5, Theorem 2.1], except that $4L^2$ gets replaced by $\frac{1}{2}(L + \alpha^m\sigma C_\chi)^2$. Notably this says that an additional step size condition is needed for our results to hold in the constant step size regime, but not necessary for a decreasing step size.

B.2 Results for SGDo

SGD with reshuffling (SGDo, also called SGD without replacement) is an optimization scheme that is also widely used in practice compared to SGDo and online SGD. In the context of CD, it corresponds to Algorithm 2 with batches chosen as

$$(B_{t,1},\ldots,B_{t,N}) \;=\pi(\{1,\ldots,n\})\,,$$

where π is a uniform draw of the permutation group on n elements. We denote the iterates of SGDo $(\psi_{t,j}^{\mathrm{SGDo}})_{t\in\mathbb{N},j\in[N]}$. Analogously to $\vartheta_{n,m,T}^{\mathrm{SGDw}}$, we define, for X_1' an i.i.d. copy of X_1, $\epsilon > 0$ and $n \in \mathbb{N}$, the tail probability term

$$\vartheta_{n,m,T}^{\mathrm{SGDo}}(\epsilon) := \sup_{\substack{t\in[T]\\ j\in[N]}} \mathbb{P}\left(\frac{\left\|\sum_{i=1}^n\left(\mathbb{E}\Big[\phi\Big(K^m_{\psi_{t-1,j}^{\mathrm{SGDo}}}(X_i)\Big)\Big|X_i,\psi_{t-1,j}^{\mathrm{SGDo}}\Big] - \mathbb{E}\Big[\phi\Big(K^m_{\psi_{t-1,j}^{\mathrm{SGDo}}}(X_1')\Big)\Big|\psi_{t-1,j}^{\mathrm{SGDo}}\Big]\right)\right\|}{n} > \epsilon\right)\,.$$

Also denote $\varepsilon_{n,m,T;\nu}^{\mathrm{SGDo}}(\epsilon) = \sqrt{\epsilon^2 + \kappa_m^2\left(\vartheta_{n,m,T}^{\mathrm{SGDo}}(\epsilon)\right)^{\frac{\nu-2}{\nu}}}$ and $\sigma_{n,T}^{\mathrm{SGDw}} = \varepsilon_{n,m,T;\nu}^{\mathrm{SGDo}}(\epsilon) + \frac{5\sigma+5\kappa_m}{\sqrt{B}}$. The following result says that $\psi_{t,j}^{\mathrm{SGDo}}$ enjoys exactly the same convergence guarantee as $\psi_{t,j}^{\mathrm{SGDo}}$ in Theorem 4.3. The statement is identical to that of Theorem B.1 and is stated in full for completeness; see Appendix F.2 for the proof, which is a slight adaptation of the proof for Theorem B.1. As before we write $\delta_{t,j}^{\mathrm{SGDo}} := \mathbb{E}\left\|\psi_{t,j}^{\mathrm{SGDo}} - \psi^*\right\|^2$.

Theorem B.2 (Convergence of CD-SGDo)**.** *Assume A1 (where Ψ may be non-compact), A2, A3, A4 and A7. Let $\eta_t = Ct^{-\beta}$ for some $\beta \in [0,1]$ and $C > 0$, and assume that $m > \frac{\log(\sigma C_\chi/\mu)}{\log|\alpha|}$ s.t. $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 0$ as in Theorem 3.2. Then for any $\epsilon > 0$, $\sqrt{\delta_{T,N}^{\mathrm{SGDo}}}$ is upper bounded by*

$$\begin{cases} E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDo}}} + C\sigma_{n,T}^{\mathrm{SGDo}}\Big(\frac{4e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}}}{\tilde{\mu}_m C} + 2N(1+\tilde{\mu}_m C)^{N-1}\varphi_{\frac{1}{2}-L^2C^2N}(T+1)\, E_2^{T,N}\Big) & \text{for } \beta = \frac{1}{2}\,, \\ E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDo}}} + C\sigma_{n,T}^{\mathrm{SGDo}}\Big(\frac{4}{\tilde{\mu}_m C} + \frac{3N\big(1+\frac{L^2C^2}{2}\big)^{N-1} e^{2L^2C^2N}\,\log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}}\Big) & \text{for } \beta = 1\,, \\ E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDo}}} + C\sigma_{n,T}^{\mathrm{SGDo}}\Big(\frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)}\frac{N}{(T+1)^\beta}} + \frac{3^\beta(1+\tilde{\mu}_m C)^{N-1}(T+2)^\beta}{L^2C^2}\, E_2^{T,N}\Big) & \text{otherwise}\,, \end{cases}$$

where $E_1^{T,N}$ and $E_2^{T,N}$ are two decreasing functions in T defined by

$$\begin{aligned} E_1^{T,N} &:= \exp\Big(1 - N\tilde{\mu}_m C\varphi_{1-\beta}(T+1) + \frac{NL^2C^2}{2}\varphi_{1-2\beta}(T+1)\Big)\,, \\ E_2^{T,N} &:= \exp\Big(-\frac{N\tilde{\mu}_m C}{2}\varphi_{1-\beta}(T+1) + 2NL^2C^2\varphi_{1-2\beta}(T+1)\Big)\,. \end{aligned}$$

Remark. We also remark that existing works [47, 48] show that the standard SGDo typically gives a faster convergence rate in T than SGDw. An analogous result for the CD setup would involve additional technical hurdles of jointly controlling the correlations across minibatches and from reusing data samples, and we defer this to future work.

B.3 Explicit tail control

We now provide the full explicit tail control bounds. All results in this section hold directly for $\varepsilon_{\nu;n,m,T}^{\mathrm{SGDo}}(\epsilon)$ and $\delta_{T,N}^{\mathrm{SGDo}}$, and we omit them here. In the result below, we denote r_Ψ as the radius of the smallest sphere in $\mathbb{R}^p$ that contains Ψ, which is finite under A1.

Lemma B.3. *Assume A5 and A6. Let $n \in \mathbb{N}$ be sufficiently large s.t. $\frac{\log n}{n} < \frac{\sigma_m^2\zeta_m^2}{p+\nu-2}$. Then*

$$\inf_{\epsilon>0}\varepsilon_{\nu;n,m,T}^{\mathrm{SGDw}}(\epsilon) \;\le\; \left(\frac{3\sigma_m\sqrt{p((\nu-2)p+2\nu)}}{\sqrt{\nu-2}} + \kappa_{\nu;m}2^{\frac{\nu-2}{2\nu}}(r_\Psi)^{\frac{(\nu-2)p}{2\nu}}\left(1 + \frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2}((\nu-2)p+2\nu)^{1/2}}\right)^{\frac{(\nu-2)p}{2\nu}}\right)\frac{\sqrt{\log n}}{\sqrt{n}}\,.$$

In particular, if we additionally assume the conditions of Theorem B.1, we have

$$\lim_{T\to\infty}\sqrt{\delta_{T,N}^{\mathrm{SGDw}}} \;\le\; C'(p,\nu,m,\Psi,\beta)\Big(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{1}{\sqrt{B}}\Big)$$

where

$$\begin{aligned} C'(p,\nu,m,\Psi,\beta) &:= \frac{8(1+5\sigma+5\kappa_m)}{\tilde{\mu}_m} \\ &\times \left(\frac{3\sigma_m\sqrt{p((\nu-2)p+2\nu)}}{\sqrt{\nu-2}} + \kappa_{\nu;m}2^{\frac{\nu-2}{2\nu}}(r_\Psi)^{\frac{(\nu-2)p}{2\nu}}\left(1 + \frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2}((\nu-2)p+2\nu)^{1/2}}\right)^{\frac{(\nu-2)p}{2\nu}}\right). \end{aligned}$$

Lemma B.4. *Assume the conditions of Theorem 4.3 and additionally that A5 holds. Then*

$$\begin{aligned} \inf_{\epsilon>0}\varepsilon_{\nu;n,m,T}^{\mathrm{SGDw}}(\epsilon) &\le \big(3C_m + \kappa_{\nu;m}^{\nu/2}(r_\Psi)^{\frac{(\nu-2)p}{2\nu}}C_m^{-\frac{\nu-2}{2}}3^{\frac{(\nu-2)p}{2\nu}}\big)\, n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}}\,, \\ \lim_{T\to\infty}\big(\delta_{T,N}^{\mathrm{SGDw}}\big)^{1/2} &\le \tilde{C}'(p,\nu,m,\Psi,\beta)\big(n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}} + B^{-1/2}\big)\,, \end{aligned}$$

where

$$\tilde{C}'(p,\nu,m,\Psi) := \frac{8(1+5\sigma+5\kappa_m)}{\tilde{\mu}_m}\big(3C_m + \kappa_{\nu;m}^{\nu/2}(r_\Psi)^{\frac{(\nu-2)p}{2\nu}}C_m^{-\frac{\nu-2}{2}}3^{\frac{(\nu-2)p}{2\nu}}\big)\,.$$

The next result considers a ϕ-discrepancy mixing condition ([35]), which is a mixing assumption on K_ψ^m but with respect to a specific test function ϕ, and we impose it uniformly over $\psi \in \Psi$. We also recall that $X_1^\psi \sim p_\psi$.

Lemma B.5. *Assume that there exist $\tilde{\alpha} \in [0,1)$ and $\tilde{C}_K > 0$ such that, for all $\psi \in \Psi$ and $x \in \mathcal{X}$, $\|\mathbb{E}[\phi(K_\psi^m(x))] - \mathbb{E}[\phi(X_1^\psi)]\| \leq \tilde{C}_K \tilde{\alpha}^m$. Then*

$$\inf_{\epsilon>0} \varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon) \;\leq\; \big(1 + 2^{\frac{\nu-2}{2\nu}} \kappa_{\nu;m} (\tilde{C}_K)^{\frac{\nu-2}{2\nu}}\big) \tilde{\alpha}^{\frac{(\nu-2)m}{3\nu-2}} \;.$$

In particular, if we additionally assume the conditions of Theorem 4.3, we have

$$\lim_{T\to\infty} \sqrt{\delta^{\mathrm{SGDw}}_{T,N}} \;\leq\; \frac{8}{\mu - \alpha^m \sigma C_\chi} \Big(\big(1 + 2^{\frac{\nu-2}{2\nu}} \kappa_{\nu;m} (\tilde{C}_K)^{\frac{\nu-2}{2\nu}}\big) \tilde{\alpha}^{\frac{(\nu-2)m}{3\nu-2}} + \frac{5\sigma + 5\kappa_m}{\sqrt{B}} \Big) \;.$$

C Auxiliary Tools

C.1 Properties of φ_γ

The following lemma collects some identities used in [22].

Lemma C.1. *φ_γ satisfies the following properties:*

(i) *φ_γ is increasing on $\mathbb{R}_+$ for all γ ;*

(ii) *$\varphi_\gamma(t) \leq \frac{t^\gamma}{\gamma}$ for $\gamma > 0$, and $\varphi_\gamma(t) \leq -\frac{1}{\gamma}$ for $\gamma < 0$;*

(iii) *$\varphi_{1-\beta}(t) \geq t^{1-\beta}$ for $\beta \in (0,1]$;*

(iv) *$\varphi_\gamma(t) - \varphi_\gamma(\frac{t}{2}) \geq \frac{1}{2} x^\gamma$ for $\gamma \in (0,1]$.*

The next lemma provides some additional results on φ_γ.

Lemma C.2. *φ_γ satisfies the following properties:*

(i) *φ_γ is positive on $t > 0$ and increasing for every $\gamma \in \mathbb{R}$;*

(ii) *For $1 \leq t_1 \leq t_2$ and $\beta \geq 0$, we have*

$$\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1) \;\leq\; \textstyle\sum_{t=t_1}^{t_2} t^{-\beta} \;\leq\; 2\big(\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)\big) \;.$$

If instead $\beta < 0$, we have

$$\frac{1}{2}\big(\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)\big) \;\leq\; \textstyle\sum_{t=t_1}^{t_2} t^{-\beta} \;\leq\; \varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1) \;;$$

(iii) *For $1 \leq t_1 \leq t_2$ and $\gamma \neq 0$, we have*

$$\begin{aligned} t_1^{\gamma-1} \;&\leq \varphi_\gamma(t_2) - \varphi_\gamma(t_1) \;\leq\; t_2^{\gamma-1} && \text{if } \gamma \geq 1 \;, \\ t_2^{-(1-\gamma)} \;&\leq \varphi_\gamma(t_2) - \varphi_\gamma(t_1) \;\leq\; t_1^{-(1-\gamma)} && \text{if } \gamma \leq 1 \;; \end{aligned}$$

(iv) *Let $1 \leq t_1 < t_2$ and $\kappa, \beta \geq 0$. If $\kappa \neq 1$ and $a > 0$, we have*

$$\textstyle\sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp\big(a\,\varphi_{1-\kappa}(t-1)\big) \;\leq\; \frac{(t_2+1)^{\max\{\kappa-\beta,0\}}}{a} \exp\big(a\,\varphi_{1-\kappa}(t_2+1)\big) \;,$$

and if $\kappa \neq 1$ and $a < 0$, we have

$$\textstyle\sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp\big(a\,\varphi_{1-\kappa}(t)\big) \;\leq\; \frac{(t_2+1)^{\max\{\kappa-\beta,0\}}}{(-a)} \exp\big(a\,\varphi_{1-\kappa}(t_1)\big) \;.$$

Proof of Lemma C.2. (i) follows from checking $\gamma > 0$, $\gamma = 0$ and $\gamma < 0$ respectively. The first set of bounds in (ii) follow by noting that $t \mapsto t^{-\beta}$ is decreasing for $\beta \geq 0$:

$$\textstyle\sum_{t=t_1}^{t_2} t^{-\beta} \;\geq\; \int_{t_1}^{t_2+1} t^{-\beta} dt \;=\; \varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1) \;,$$

$$\textstyle\sum_{t=t_1}^{t_2} t^{-\beta} \;\leq 2 \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \;\leq\; 2 \int_{t_1-1}^{t_2} (t+1)^{-\beta} dt \;=\; 2\big(\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)\big) \;.$$

The second set of bounds follows from noting that $t \mapsto t^{-\beta}$ is increasing for $\beta < 0$:

$$\sum_{t=t_1}^{t_2} t^{-\beta} \;\geq\; \frac{1}{2}\sum_{t=t_1}^{t_2}(t+1)^{-\beta} \;\geq\; \frac{1}{2}\int_{t_1-1}^{t_2}(t+1)^{-\beta}dt \;=\; \frac{1}{2}\big(\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)\big)\;,$$

$$\sum_{t=t_1}^{t_2} t^{-\beta} \;\leq\; \int_{t_1}^{t_2+1} t^{-\beta}dt \;=\; \varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)\;.$$

For (iii), we note that for $\gamma \neq 0$,

$$\varphi_\gamma(t_2) - \varphi_\gamma(t_1) \;=\; \frac{t_2^\gamma - t_1^\gamma}{\gamma}\;,$$

so by the mean value theorem,

$$\inf_{t_1 \leq t \leq t_2} t^{\gamma-1} \;\leq\; \varphi_\gamma(t_2) - \varphi_\gamma(t_1) \;\leq\; \sup_{t_1 \leq t \leq t_2} t^{\gamma-1}\;.$$

The desired bounds then follow from an explicit computation of the infimum and the maximum in each of the two cases $\gamma \geq 1$ and $\gamma \leq 1$.

For (iv), we first consider the case $\kappa \neq 1$ and $a > 0$. Then

$$\begin{aligned}
\sum_{t=t_1}^{t_2}(t+1)^{-\beta}\exp\big(a\,\varphi_{1-\kappa}(t-1)\big) \;&=\; e^{-\frac{a}{1-\kappa}}\sum_{t=t_1}^{t_2}(t+1)^{-\beta}\exp\left(\frac{at^{1-\kappa}}{1-\kappa}\right)\\
&\leq e^{-\frac{a}{1-\kappa}}\max_{t_1\leq t\leq t_2}\left(\frac{(t+1)^\kappa}{(t+1)^\beta}\right)\sum_{t=t_1}^{t_2}(t+1)^{-\kappa}\exp\left(\frac{at^{1-\kappa}}{1-\kappa}\right)\\
&\leq e^{-\frac{a}{1-\kappa}}(t_2+1)^{\max\{\kappa-\beta,0\}}\sum_{t=t_1}^{t_2}(t+1)^{-\kappa}\exp\left(\frac{at^{1-\kappa}}{1-\kappa}\right)\;.
\end{aligned}$$

Since, for $x \geq 0$, $x \mapsto (x+1)^{-\kappa}$ is decreasing and $x \mapsto \exp(ax^{1-\kappa}/(1-\kappa))$ is increasing, we have that for $t_1 \leq t \leq t_2$ and $x \in [t, t+1]$,

$$(t+1)^{-\kappa} \;\leq x^{-\kappa} \qquad \text{and} \qquad \exp(at^{1-\kappa}/(1-\kappa)) \;\leq\; \exp(ax^{1-\kappa}/(1-\kappa))\;.$$

This implies that

$$\begin{aligned}
&\sum_{t=t_1}^{t_2}(t+1)^{-\beta}\exp\big(a\,\varphi_{1-\kappa}(t-1)\big)\\
&\qquad\leq (t_2+1)^{\max\{\kappa-\beta,0\}}e^{-\frac{a}{1-\kappa}}\sum_{t=t_1}^{t_2}\int_t^{t+1} x^{-\kappa}\exp\left(\frac{ax^{1-\kappa}}{1-\kappa}\right)dx\\
&\qquad= (t_2+1)^{\max\{\kappa-\beta,0\}}e^{-\frac{a}{1-\kappa}}\int_{t_1}^{t_2+1} x^{-\kappa}\exp\left(\frac{ax^{1-\kappa}}{1-\kappa}\right)dx\\
&\qquad\leq \frac{(t_2+1)^{\max\{\kappa-\beta,0\}}}{a}\,e^{-\frac{a}{1-\kappa}}\,e^{\frac{a(t_2+1)^{1-\kappa}}{1-\kappa}}\\
&\qquad= \frac{(t_2+1)^{\max\{\kappa-\beta,0\}}}{a}\,\exp\big(a\,\varphi_{1-\kappa}(t_2+1)\big)\;.
\end{aligned}$$

The main difference in the case $\kappa \neq 1$ and $a < 0$ is that we now use $x \mapsto \exp(a(x+1)^{1-\kappa}/(1-\kappa))$ is decreasing to obtain, for $t_1 \leq t \leq t_2$ and $x \in [t, t+1]$,

$$\exp(a(t+1)^{1-\kappa}/(1-\kappa)) \;\leq\; \exp(ax^{1-\kappa}/(1-\kappa))\;.$$

A similar argument then yields

$$\begin{aligned}
&\sum_{t=t_1}^{t_2}(t+1)^{-\beta}\exp\big(a\,\varphi_{1-\kappa}(t)\big)\\
&\leq (t_2+1)^{\max\{\kappa-\beta,0\}}e^{-\frac{a}{1-\kappa}}\sum_{t=t_1}^{t_2}(t+1)^{-\kappa}\exp\left(\frac{a(t+1)^{1-\kappa}}{1-\kappa}\right)\\
&\leq (t_2+1)^{\max\{\kappa-\beta,0\}}e^{-\frac{a}{1-\kappa}}\int_{t_1}^{t_2+1} x^{-\kappa}\exp\left(\frac{ax^{1-\kappa}}{1-\kappa}\right)dx\\
&= \frac{(t_2+1)^{\max\{\kappa-\beta,0\}}}{a}\Big(\exp\big(a\,\varphi_{1-\kappa}(t_2+1)\big) - \exp\big(a\,\varphi_{1-\kappa}(t_1)\big)\Big)\\
&\leq \frac{(t_2+1)^{\max\{\kappa-\beta,0\}}}{(-a)}\,\exp\big(a\,\varphi_{1-\kappa}(t_1)\big)\;.
\end{aligned}$$

□

We also need the following lemma, which is useful for controlling the accumulation of errors from the noise terms over iterations.

Lemma C.3. *For any $a, b \geq 0$, $T, N \in \mathbb{N}$ and $\kappa, \beta \geq 0$ such that $bt^{-\beta} - at^{-\kappa} \leq 1$ for all $1 \leq t \leq T$, we have that*

$$\prod_{t=1}^{T}(1 - bt^{-\beta} + at^{-\kappa})^N \;\leq\; \exp\big(- bN\,\varphi_{1-\beta}(T+1) + aN\,\varphi_{1-\kappa}(T+1)\big)\;.$$

Moreover, for any $\zeta \geq 0$, we have that

$$\sum_{t=1}^{T} t^{-\zeta}\left(\sum_{j=1}^{N}(1 - bt^{-\beta} + at^{-\kappa})\right)\prod_{s=t+1}^{T}(1 - bs^{-\beta} + as^{-\kappa})^N \\ \leq Q_1 + \exp\Big(- \frac{bN}{2}\,\varphi_{1-\beta}(T+1) + 4aN\,\varphi_{1-\kappa}(T+1)\Big)\,Q_2\;,$$

where

$$Q_1 \;\coloneqq\; \begin{cases} \dfrac{2^{2\zeta+1}(T+3)^{\max\{\beta-\zeta,0\}}}{b}\exp\Big(\dfrac{bN}{2(1-\beta)(T+1)^{\beta}}\Big) & \text{if } \beta \neq 1\,,\; b > 0\;, \\ 2N\varphi_{1-\zeta+bN/2}(T+1)\exp\Big(- \dfrac{bN}{2}\,\varphi_{1-\beta}(T+1)\Big) & \text{if } \beta = 1 \text{ or } b = 0\;, \end{cases}$$

and

$$Q_2 \;\coloneqq\; \begin{cases} \dfrac{3^{\zeta}(1+a)^{N-1}}{2a}(T+2)^{\max\{\kappa-\zeta,0\}} & \text{if } \kappa \neq 1 \text{ and } a > 0\;, \\ 2N(1+a)^{N-1}\,\varphi_{1-\zeta-2aN}(T+1) & \text{if } \kappa = 1 \text{ or } a = 0\;. \end{cases}$$

In the special case $\zeta = \beta = 1 < \kappa$, we have

$$\sum_{t=1}^{T} t^{-\zeta}\left(\sum_{j=1}^{N}(1 - bt^{-\beta} + at^{-\kappa})^{j-1}\right)\prod_{s=t+1}^{T}(1 - bs^{-\beta} + as^{-\kappa})^N \\ \leq\; \frac{4}{b} + \frac{3N(1+a)^{N-1}\,e^{\frac{4aN}{\kappa-1}}\,\log(T+1)}{(T+1)^{\frac{bN}{2}}}\;.$$

Proof of Lemma C.3. By assumption, $bt^{-\beta} - at^{-\kappa} \leq 1$ for all $1 \leq t \leq T$. Since $0 \leq 1 - x \leq e^{-x}$ for all $x \leq 1$, we have that for any $1 \leq t_1 \leq t_2 \leq T$,

$$\prod_{t=t_1}^{t_2}(1 - bt^{-\beta} + at^{-\kappa})^N \;\leq\; \exp\left(- bN\sum_{t=t_1}^{t_2} t^{-\beta} + aN\sum_{t=t_1}^{t_2} t^{-\kappa}\right)\;. \tag{13}$$

Applying this to the first quantity of interest followed by noting that $a, b \geq 0$ and using Lemma C.2(ii), we obtain the first bound that

$$\prod_{t=1}^{T}(1 - bt^{-\beta} + at^{-\kappa})^N \;\leq\; \exp\left(- bN\sum_{t=1}^{T} t^{-\beta} + aN\sum_{t=1}^{T} t^{-\kappa}\right) \\ \leq\; \exp\big(- bN\,\varphi_{1-\beta}(T+1) + aN\,\varphi_{1-\kappa}(T+1)\big)\;.$$

For the second bound, we define

$$t_0 \;\coloneqq\; \sup\left\{ t \leq T \;\middle|\; \frac{b}{2} \leq at^{-(\kappa-\beta)} \right\}\;.$$

Then by noting that $1 - bt^{-\beta} + at^{-\kappa} \geq 0$ for all $1 \leq t \leq T$ again,we can bound the quantity of interest as

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{T} t^{-\zeta} \left(\sum_{j=1}^{N} (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{T} (1 - bs^{-\beta} + as^{-\kappa})^{N} \\
&= \textstyle\sum_{t=t_0+1}^{T} t^{-\zeta} \left(\sum_{j=1}^{N} (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{T} (1 - bs^{-\beta} + as^{-\kappa})^{N} \\
&\quad + \left(\textstyle\prod_{s=t_0+1}^{T} (1 - bs^{-\beta} + as^{-\kappa}) \right) \\
&\qquad \times \left(\textstyle\sum_{t=1}^{t_0} t^{-\zeta} \left(\sum_{j=1}^{N} (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{t_0} (1 - bs^{-\beta} + as^{-\kappa})^{N} \right) \\
&\leq \textstyle\sum_{t=t_0+1}^{T} t^{-\zeta} \left(\sum_{j=1}^{N} \left(1 - \frac{b}{2} t^{-\beta}\right)^{j-1} \right) \prod_{s=t+1}^{T} \left(1 - \frac{b}{2} s^{-\beta}\right)^{N} \\
&\quad + \left(\textstyle\prod_{s=t_0+1}^{T} \left(1 - \frac{b}{2} s^{-\beta}\right)^{N} \right) \left(\textstyle\sum_{t=1}^{t_0} t^{-\zeta} \left(\sum_{j=1}^{N} (1 + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{t_0} (1 + as^{-\kappa})^{N} \right) \\
&\leq N \times \underbrace{\textstyle\sum_{t=t_0+1}^{T} t^{-\zeta} \prod_{s=t+1}^{T} \left(1 - \frac{b}{2} s^{-\beta}\right)^{N}}_{=:S_1} \\
&\quad + N(1+a)^{N-1} \times \underbrace{\left(\textstyle\prod_{s=t_0+1}^{T} \left(1 - \frac{b}{2} s^{-\beta}\right)^{N} \right)}_{=:S_3} \times \underbrace{\left(\textstyle\sum_{t=1}^{t_0} t^{-\zeta} \prod_{s=t+1}^{t_0} (1 + as^{-\kappa})^{N} \right)}_{=:S_2} .
\end{aligned}
$$

In the last line, we have used that $0 \leq 1 - \frac{b}{2} t^{-\beta} \leq 1$ for $t \geq t_0 + 1$ and $1 + at^{-\kappa} \leq 1 + a$. To control the three quantities, we first note that by (13) , we have

$$
\begin{aligned}
S_3 &\leq \exp\left(-\frac{bN}{2} \textstyle\sum_{s=1}^{T} s^{-\beta} \right) \exp\left(\frac{bN}{2} \textstyle\sum_{s=1}^{t_0} s^{-\beta} \right) \\
&\overset{(a)}{\leq} \exp\left(-\frac{bN}{2} \textstyle\sum_{s=1}^{T} s^{-\beta} \right) \exp\left(aN \textstyle\sum_{s=1}^{t_0} s^{-\kappa} \right) \\
&\overset{(b)}{\leq} \exp\left(-\frac{bN}{2} \varphi_{1-\beta}(T+1) + 2aN \varphi_{1-\kappa}(T+1) \right) .
\end{aligned}
$$

In (a) above, we have noted that $\frac{b}{2} \leq as^{-(\kappa-\beta)}$ for $s \leq t_0$; in (b), we have used $t_0 \leq T$ and Lemma C.2(ii) with $a, b \geq 0$. In the special case $\beta = 1 < \kappa$, the above yields

$$
\begin{aligned}
S_3 &\leq (T+1)^{-\frac{bN}{2}} \exp\left(2aN \frac{1 - (T+1)^{-(\kappa-1)}}{\kappa - 1} \right) \\
&\leq (T+1)^{-\frac{bN}{2}} e^{\frac{2aN}{\kappa-1}} .
\end{aligned}
$$

We now control S_2. By (13) again, we have

$$
\begin{aligned}
S_2 &\leq \textstyle\sum_{t=1}^{t_0} t^{-\zeta} \exp\left(aN \textstyle\sum_{s=t+1}^{t_0} s^{-\kappa} \right) \\
&\leq \textstyle\sum_{t=1}^{T} t^{-\zeta} \exp\left(aN \textstyle\sum_{s=t+1}^{T} s^{-\kappa} \right) \\
&\overset{(c)}{\leq} \exp\left(2aN\varphi_{1-\kappa}(T+1) \right) \times \left(\textstyle\sum_{t=1}^{T} t^{-\zeta} \exp\left(-2aN\varphi_{1-\kappa}(t+1) \right) \right) \\
&\overset{(d)}{\leq} 3^{\zeta} \exp\left(2aN\varphi_{1-\kappa}(T+1) \right) \textstyle\sum_{t=1}^{T} (t+2)^{-\zeta} \exp\left(-2aN\varphi_{1-\kappa}(t+1) \right) .
\end{aligned}
$$

In (c) above, we have applied Lemma C.2(ii); in (d), we have noted that $\sup_{t \in \mathbb{N}} (t+2)^{\beta}/t^{\beta} = 3^{\beta}$. If $\kappa \neq 1$ and $a > 0$, we can apply Lemma C.2(iv) to get that

$$
\begin{aligned}
S_2 &\leq \frac{3^{\zeta}}{2aN} (T+2)^{\max\{\kappa-\zeta, 0\}} \exp\left(2aN\varphi_{1-\kappa}(T+1) \right) \\
&= \frac{Q_2}{N(1+a+c)^{N-1}} \exp\left(2aN\varphi_{1-\kappa}(T+1) \right) .
\end{aligned}
$$

If $\kappa = 1$ or $a = 0$, the bound from (c) above reads

$$\begin{aligned} S_2 &\le \exp\big(2aN\varphi_{1-\kappa}(T+1)\big) \textstyle\sum_{t=1}^T t^{-\zeta}(t+1)^{-2aN} \\ &\le \exp\big(2aN\varphi_{1-\kappa}(T+1)\big) \textstyle\sum_{t=1}^T t^{-\zeta-2aN} \\ &\le 2\,\varphi_{1-\zeta-2aN}(T+1)\,\exp\big(2aN\varphi_{1-\kappa}(T+1)\big) \;=\; \frac{Q_2}{N(1+a)^{N-1}}\,\exp\big(2aN\varphi_{1-\kappa}(T+1)\big)\;, \end{aligned}$$

where we have used Lemma C.2(ii) in the last line. Now consider the special case with $\zeta = 1 < \kappa$, the bound from (d) becomes

$$\begin{aligned} S_2 &\le 3\exp\big(2aN\varphi_{1-\kappa}(T+1)\big)\Big(\textstyle\sum_{t=1}^T (t+2)^{-1}\exp\big(-2aN\varphi_{1-\kappa}(t+1)\big)\Big) \\ &\le 3\exp\Big(2aN\,\frac{1-(T+1)^{-(\kappa-1)}}{\kappa-1}\Big)\textstyle\sum_{t=1}^T (t+2)^{-1} \\ &\le 3e^{\frac{2aN}{\kappa-1}}\log(T+1)\;. \end{aligned}$$

We are left with controlling S_1, which follows from a similar strategy as controlling S_2:

$$\begin{aligned} S_1 &\overset{(13)}{\le} \textstyle\sum_{t=t_0+1}^T t^{-\zeta}\exp\Big(-\frac{bN}{2}\sum_{s=t+1}^T s^{-\beta}\Big) \\ &\le \textstyle\sum_{t=1}^T t^{-\zeta}\exp\Big(-\frac{bN}{2}\sum_{s=t+1}^T s^{-\beta}\Big) \\ &\overset{(a)}{\le} \exp\Big(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\Big)\textstyle\sum_{t=1}^T t^{-\zeta}\exp\Big(\frac{bN}{2}\varphi_{1-\beta}(t+1)\Big) \\ &\le 4^{\zeta}\exp\Big(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\Big)\textstyle\sum_{t=1}^T (t+3)^{-\zeta}\exp\Big(\frac{bN}{2}\varphi_{1-\beta}(t+1)\Big)\;. \end{aligned} \tag{14}$$

In (a) above, we used Lemma C.2(ii). For $\beta \ne 1$ and $b \ne 0$, we can apply Lemma C.2(iv) with $\frac{b}{2} > 0$ to obtain

$$\begin{aligned} S_1 &\le 4^{\zeta}\exp\Big(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\Big)\frac{(T+3)^{\max\{\beta-\zeta,0\}}}{bN/2}\exp\Big(\frac{bN}{2}\varphi_{1-\beta}(T+3)\Big) \\ &= \frac{2^{2\zeta+1}(T+3)^{\max\{\beta-\zeta,0\}}}{bN}\exp\Big(\frac{bN}{2(1-\beta)}\Big((T+3)^{1-\beta}-(T+1)^{1-\beta}\Big)\Big) \\ &\overset{(b)}{\le} \frac{2^{2\zeta+1}(T+3)^{\max\{\beta-\zeta,0\}}}{bN}\exp\Big(\frac{bN}{2(1-\beta)(T+1)^{\beta}}\Big) \;=\; \frac{Q_1}{N}\;. \end{aligned}$$

In (b), we have used Lemma C.2(iii) with $1-\beta \le 1$. Meanwhile, if $\beta = 1$ or $b = 0$, we have

$$\begin{aligned} S_1 &\le \exp\Big(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\Big)\textstyle\sum_{t=1}^T t^{-\zeta}(t+1)^{bN/2} \\ &\le \exp\Big(-\frac{b}{2}\varphi_{1-\beta}(T+1)\Big)\textstyle\sum_{t=1}^T t^{-\zeta+bN/2} \\ &\le 2\varphi_{1-\zeta+bN/2}(T+1)\exp\Big(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\Big) \;=\; \frac{Q_1}{N}\;. \end{aligned}$$

For the special case with $\zeta = \beta = 1$, the bound in (14) becomes

$$\begin{aligned} S_1 &\le \exp\Big(-\frac{bN}{2}\varphi_0(T+1)\Big)\textstyle\sum_{t=1}^T t^{-1}\exp\Big(\frac{bN}{2}\varphi_0(t+1)\Big) \\ &= (T+1)^{-\frac{bN}{2}}\textstyle\sum_{t=1}^T t^{-1}(t+1)^{\frac{bN}{2}} \\ &\le (T+1)^{-\frac{bN}{2}}\textstyle\sum_{t=1}^T t^{-(1-\frac{bN}{2})} \\ &\overset{(c)}{\le} 2(T+1)^{-\frac{bN}{2}}\varphi_{\frac{bN}{2}}(T+1) \;=\; 2(T+1)^{-\frac{bN}{2}}\,\frac{(T+1)^{bN/2}-1}{bN/2} \;\le\; \frac{4}{bN}\;. \end{aligned}$$

In (c), we have used Lemma C.2(ii) for both the case $1-\frac{bN}{2} \le 0$ and $1-\frac{bN}{2} \ge 0$.

Combining the bounds for the general cases, we obtain the first desired inequality that

$$\begin{aligned} &\textstyle\sum_{t=1}^T t^{-\zeta}\Big(\sum_{j=1}^N (1-bt^{-\beta}+at^{-\kappa})^{j-1}\Big)\prod_{s=t+1}^T (1-bs^{-\beta}+as^{-\kappa})^N \\ &\le\; Q_1 + \exp\Big(-\frac{b}{2}\varphi_{1-\beta}(T+1)+u\varphi_{1-\xi}(T+3)+4a\,\varphi_{1-\kappa}(T+1)\Big)Q_2\;. \end{aligned}$$

For the special case $\zeta = \beta = 1 < \kappa, \gamma$, combining the earlier bounds gives

$$\sum_{t=1}^{T} t^{-\zeta} \left(\sum_{j=1}^{N} (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{T} (1 - bs^{-\beta} + as^{-\kappa})^{N} \leq \frac{4}{b} + \frac{3N(1+a)^{N-1} e^{\frac{4aN}{\kappa-1}} \log(T+1)}{(T+1)^{\frac{bN}{2}}} .$$

□

C.2 Contraction and integrability results

The next result is a standard result in convex analysis, needed to handle projections performed in Algorithms 1 and 2.

Lemma C.4. *Let Ψ a be convex subset of $\mathbb{R}^p$. Let $\psi^\star \in \Psi$. Then, for all $\psi \in \mathbb{R}^p$, we have:*

$$\|\mathrm{Proj}_\Psi(\psi) - \psi^\star\| \leq \|\psi - \psi^\star\|$$

Proof. We have:

$$\begin{aligned} \|\psi - \psi^\star\|^2 &= \|\psi - \mathrm{Proj}_\Psi(\psi) + \mathrm{Proj}_\Psi(\psi) - \psi^\star\|^2 \\ &= \|\psi - \mathrm{Proj}_\Psi(\psi)\|^2 + 2\langle \psi - \mathrm{Proj}_\Psi(\psi), \mathrm{Proj}_\Psi(\psi) - \psi^\star \rangle + \|\mathrm{Proj}_\Psi(\psi) - \psi^\star\|^2 \end{aligned}$$

Since by [49, Proposition 1.1.9], we have:

$$\langle \psi - \mathrm{Proj}_\Psi(\psi), \psi' - \mathrm{Proj}_\Psi(\psi) \rangle \leq 0$$

for all $\psi' \in \Psi$, we can use this inequality at $\psi' = \psi^\star \in \Psi$ to obtain:

$$\|\psi - \psi^\star\|^2 \geq \|\psi - \mathrm{Proj}_\Psi(\psi)\|^2$$

and the result follows by taking the square root. □

We now state two lemmas that guarantee an amount of integrability sufficient to our analysis.

Lemma C.5. *Let $p, q \in \mathcal{P}(\mathcal{X})$ such that $\frac{\mathrm{d}p}{\mathrm{d}q}$ exists, and such that $\chi^2(p;q) < +\infty$. Assume that $f \in L^2(q)$ Then $|\mathbb{E}_p f| < +\infty$.*

Proof. By assumption, $f \in L^2(q)$. Moreover, $\chi^2(p,q) < +\infty$, and thus we have $\frac{\mathrm{d}p}{\mathrm{d}q} - 1 \in L^2(q)$. Thus, the inner product is finite, and

$$\begin{aligned} &\left| \int f \left(\frac{\mathrm{d}p}{\mathrm{d}q} - 1 \right) \mathrm{d}q \right| = \left| \int f \mathrm{d}p - \int f \mathrm{d}q \right| = |\mathbb{E}_p f - \mathbb{E}_q f| \coloneqq M < +\infty \\ &\implies M - |\mathbb{E}_q f| < \mathbb{E}_p f < M + |\mathbb{E}_q f| \end{aligned}$$

□

Lemma C.6. *For all $\psi \in \Psi$, for all $m \geq 1$, and for all $k \geq 1$, we have:*

$$\mathbb{E}_{p_{\psi^\star}} \left\| P_\psi^m \phi \right\|^k < +\infty$$

Proof. By analycity of the log partition function $\psi \longmapsto \log Z(\psi)$, we have $\int \|\phi\|^k \mathrm{d}p_\psi(x) < +\infty$ for all $\psi \in \Psi$, and thus, the function $x \longmapsto \|\phi\|^k (x) \in L^2(p_\psi)$ for all ψ. Consequently, $P_\psi^m \|\phi\|^k \in L^2(p_\psi)$. We can apply Lemma C.5 to $P_p si^m \|\phi\|^k$ to obtain $\mathbb{E}_{\psi^\star} P^m \|\phi\|^k < +\infty$ for all $k \geq 1$ and for all $m \geq 0$. As a by-product, we obtain $P_\psi^m \|\phi\|^k \in L^2(p_{\psi^\star})$, and thus so $\|P^m \phi\|^k$. □

The following lemma is used multiple time in our analysis.

Lemma C.7. *Assume A3. Then, for any $f \in \{\phi_i\}_{i=1}^p \cup \{\phi_i \phi_j\}_{i,j=1}^p$ and $\psi \in \Psi$, we have*

$$\left\| \mathbb{E}_{p_{\psi^\star}} \left[P_\psi^m (f - \mathbb{E}_{p_\psi} f) \right] \right\| \leq \alpha^m C_\chi \left(\mathbb{E}_{p_\psi} \left[\left\| f - \mathbb{E}_{p_\psi} f \right\|^2 \right] \right)^{1/2} \|\psi - \psi^\star\|$$

Proof. Let us note first that

$$
\begin{aligned}
\|\mathbb{E}_{p_{\psi^\star}} P_\psi^m \left(f - \mathbb{E}_{p_\psi} f\right)\|^2 &\overset{(a)}{=} \sum_{i=1}^p \left(\int P_\psi^m \left(f_i - \mathbb{E}_{p_\psi} f_i\right)(x) \left(p_{\psi^\star}(x) - p_\psi(x)\right) \mathrm{d}x \right)^2 \\
&= \sum_{i=1}^p \left(\int P_\psi^m \left(f_i - \mathbb{E}_{p_\psi} f_i\right)(x) \left(\frac{\mathrm{d}p_{\psi^\star}}{\mathrm{d}p_\psi}(x) - 1 \right) p_\psi(x) \mathrm{d}x \right)^2 \\
&\overset{(b)}{\leq} \left(\int \left(\frac{\mathrm{d}p_\psi^\star}{\mathrm{d}p_\psi}(x) - 1 \right)^2 p_\psi(x) \mathrm{d}x \right) \sum_{i=1}^p \int P_\psi^m \left(f_i - \mathbb{E}_{p_\psi} f_i\right)(x)^2 p_\psi(x) \mathrm{d}x \\
&\leq \chi^2(p_\psi, p_{\psi^\star}) \sum_{i=1}^p \left\| P_\psi^m (f_i - \mathbb{E}_{p_\psi} f_i) \right\|_{L^2_{p_\psi}(\mathbb{R}^d)} \\
&\overset{(c)}{\leq} \alpha^{2m} \chi^2(p_\psi, p_{\psi^\star}) \sum_{i=1}^p \left\| f_i - \mathbb{E}_{p_\psi} f_i \right\|_{L^2_{p_\psi}(\mathbb{R}^d)} \\
&\overset{(d)}{\leq} \alpha^{2m} C_\chi^2 \|\psi - \psi^\star\|^2 \, \mathbb{E}_{p_\psi} \left\| f - \mathbb{E}_{p_\psi} f \right\|^2 .
\end{aligned}
$$

Here, we used the fact that P_ψ^m admits p_ψ as an invariant measure in (a) [34, Eq. (1.2.2)], the Cauchy-Schwarz inequality in (b) □

C.3 Miscellaneous

Lemma C.8. *Let $f : \Psi \to \mathbb{R}^p$ be a differentiable function in the interior of $\Psi \subseteq \mathbb{R}^p$. For $\psi \in \Psi$, define $\sigma_{\min}(\psi) \coloneqq \inf_{\theta \in \Psi, \|\theta\|=1} \theta^\top \nabla f(\psi) \theta$ and $\sigma_{\max}(\psi) \coloneqq \sup_{\theta \in \Psi, \|\theta\|=1} \theta^\top \nabla f(\psi) \theta$ with respect to the Jacobian matrix $\nabla f(\psi)$. Then for any $\psi_1, \psi_2 \in \Psi$, we have that*

$$
\inf_{\psi \in \Psi} \sigma_{\min}(\psi) \;\leq\; (\psi_1 - \psi_2)^\top \big(f(\psi_1) - f(\psi_2)\big) \;\leq\; \sup_{\psi \in \Psi} \sigma_{\max}(\psi)
$$

Proof of Lemma C.8. By the mean value theorem, there exists some $a \in (0, 1)$ such that

$$
\begin{aligned}
(\psi_1 - \psi_2)^\top \big(f(\psi_1) - f(\psi_2)\big) &= (\psi_1 - \psi_2)^\top \big(f(1 \times \psi_1 + 0 \times \psi_2) - f(0 \times \psi_1 + 0 \times \psi_2)\big) \\
&= (\psi_1 - \psi_2)^\top \nabla f(a\psi_1 + (1-a)\psi_2)(\psi_1 - \psi_2) \\
&= \|\psi_1 - \psi_2\|^2 \, \frac{(\psi_1 - \psi_2)^\top}{\|\psi_1 - \psi_2\|} \, \nabla f(a\psi_1 + (1-a)\psi_2) \, \frac{\psi_1 - \psi_2}{\|\psi_1 - \psi_2\|} .
\end{aligned}
$$

Plugging in the definition of $\sigma_{\max}$ gives the desired upper bound and similarly $\sigma_{\min}$ implies the lower bound. □

D Proofs for Online CD

D.1 Auxiliary Lemmas for Online CD

We recall the following notations used in the next lemmas, namely $\sigma_\psi \coloneqq \mathbb{E}_{p_\psi} \|\phi - \mathbb{E}_{p_\psi} \phi\|^2$, $\sigma_\star \coloneqq \sigma_{\psi^\star}$ and $\sigma \coloneqq \sup_{\psi \in \Psi} \sigma_\psi$.

We now provide two intermediary lemmas necessary to analyze the impact of variance in the CD gradient. The strategy is similar in both of them: we change the integration from $p_{\psi^\star}$ to p_ψ to obtain contraction, at the cost of an additional term scaling with $C_\chi \|\psi - \psi^\star\|$. We obtain exact constants that we choose to describe in terms of the smoothness parameters of the problem, e.g. the kth derivatives of the log partition function $\log Z$, which, for $k \geq 2$, equals the kth derivative of the negative cross-entropy model w.r.t $p_{\psi^\star}$.

Second Moment convergence The following lemmas guarantee the second moment of a sample from $k_\psi^m p_{\psi^\star}$ approaches the second moment of a sample from the target distribution p_ψ.

Lemma D.1. *Under A1, A2 and A3, for all $\psi \in \Psi$, we have:*

$$
\left| \mathbb{E}_{p_{\psi^\star}} P_\psi^m \|\phi\|^2 - \mathbb{E}_{p_\psi} \|\phi\|^2 \right| \leq \alpha^m C_\chi \|\psi - \psi^\star\| \, \|\log Z\|_{1,\infty}
$$

where

$$\|\log Z\|_{1,\infty} := \sup_{\psi\in\Psi} \sum_{i=1}^{p} (4\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + 2\partial_i^2 \log Z(\psi)^2 + 4\partial_i^1 \log Z(\psi)\partial_i^3 \log Z(\psi) + \partial_i^4 \log Z(\psi))^{1/2} < +\infty$$

Proof. Applying Lemma C.7 to each $f_i := \phi_i^2$, we have

$$\begin{aligned}\left|\mathbb{E}_{p_{\psi^\star}} P_\psi^m \phi_i^2 - \mathbb{E}_{p_\psi} \phi_i^2\right| &= \alpha^m C_\chi \|\psi - \psi^\star\| \left(\mathbb{E}_{p_\psi} \left(\phi_i^2 - \mathbb{E}_{p_\psi} \phi_i^2\right)^2\right)^{1/2} \\ &= \alpha^m C_\chi \|\psi - \psi^\star\| \left(\mathbb{E}_{p_\psi} \phi_i^4 - (\mathbb{E}_{p_\psi} \phi_i^2)^2\right)^{1/2}\end{aligned}$$

We map the two moments to derivatives of $\log Z(\psi)$, since the kth derivative of $\log Z(\psi)$ is the kth cumulant. It can be shown, using the multivariate moment to cumulant mapping, that

$$\begin{aligned}\mathbb{E}\phi_i^4 &= \frac{\partial \log Z}{\partial \psi_i}^4 + 6\frac{\partial \log Z}{\partial \psi_i}^2 \frac{\partial^2 \log Z}{\partial \psi_i^2} + 3\left(\frac{\partial^2 \log Z}{\partial \psi_i^2}\right)^2 + 4\frac{\partial \log Z}{\partial \psi_i}\frac{\partial^3 \log Z}{\partial \psi_i^3} + \frac{\partial^4 \log Z}{\partial \psi_i^4} \\ &= \partial_i^1 \log Z(\psi)^4 + 6\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + 3\partial_i^2 \log Z(\psi)^2 + 4\partial_i^1 \log Z(\psi)\partial_i^3 \log Z(\psi) \\ &\quad + \partial_i^4 \log Z(\psi)\end{aligned}$$

where $\partial_i^k \log Z(\psi)$ denotes the kth derivative of $\log Z$ with respect to ψ_i. On the other hand,

$$\begin{aligned}\mathbb{E}\phi_i^2 &= \partial_i^1 \log Z(\psi)^2 + \partial_i^2 \log Z(\psi) \\ \implies (\mathbb{E}\phi_i^2)^2 &= \partial_i^1 \log Z(\psi)^4 + 2\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + \partial_i^2 \log Z(\psi)^2\end{aligned}$$

implying

$$\begin{aligned}\mathbb{E}_{p_\psi} \phi_i^4 - (\mathbb{E}_{p_\psi^2} \phi_i^2)^2 &= 4\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + 2\partial_i^2 \log Z(\psi)^2 + 4\partial_i^1 \log Z(\psi)\partial_i^3 \log Z(\psi) \\ &\quad + \partial_i^4 \log Z(\psi)\end{aligned}$$

The result follows by summing over i, since:

$$|\mathbb{E}_{p_{\psi^\star}} P_\psi^m \|\phi\|^2 - \mathbb{E}_{p_\psi} \|\phi\|^2| \;\le\; \textstyle\sum_{i=1}^d |\mathbb{E}_{p_{\psi^\star}} P_\psi^m \phi_i^2 - \mathbb{E}_{p_\psi} \phi_i^2|$$

Note that $\|\log Z\|_{1,\infty}$ is finite since Ψ is compact and $\log Z$ is analytic. □

Squared First Moment convergence The next lemma provides convergence (in squared absolute value) of the first moment of the m-iterated Markov kernel k_ψ^m.

Lemma D.2. *Under A1, A2 and A3, for all $\psi \in \Psi$, we have*

$$\left|\mathbb{E}_{p_{\psi^\star}} \|P_\psi^m \phi\|^2 - \left\|\mathbb{E}_{p_\psi} \phi\right\|^2\right| \le \alpha^m \sigma_\psi^2 + C_\chi \alpha^{m/4} \|\log Z\|_{2,\infty} \|\psi - \psi^\star\|$$

where

$$\|\log Z\|_{2,\infty} := \sup_{\psi\in\Psi} \textstyle\sum_{i=1}^p \left(F(\psi)\partial_i^2 \log Z(\psi)\right)^{1/4} + 2\left|\partial_i^1 \log Z(\psi)\partial_i^2 \log Z(\psi)^{1/2}\right|$$

and

$$F(\psi) := 15\partial_i^2 \log Z(\psi)^3 + 10\partial_i^3 \log Z(\psi)^2 + 15\partial_i^2 \log Z(\psi)\partial_i^4 \log Z(\psi) + \partial_i^6 \log Z(\psi)$$

Proof. We have:

$$\begin{aligned}\mathbb{E}_{p_{\psi^\star}} (P_\psi^m \phi_i)^2 &= \mathbb{E}_{p_{\psi^\star}} (P_\psi^m \phi_i - \mathbb{E}_{p_\psi} \phi_i + \mathbb{E}_{p_\psi} \phi_i)^2 \\ &= \mathbb{E}_{p_{\psi^\star}} (P^m \phi_i - \mathbb{E}_{p_\psi} \phi_i)^2 + 2\mathbb{E}_{p_\psi} \phi_i \, \mathbb{E}_{p_{\psi^\star}} P^m (\phi_i - \mathbb{E}_{p_\psi} \phi_i) \\ &\quad + (\mathbb{E}_\psi \phi_i)^2 \\ \implies |\mathbb{E}_{p_{\psi^\star}} (P_\psi^m \phi_i)^2 - (\mathbb{E}_{p_\psi} \phi_i)^2| &\le \underbrace{\left|\mathbb{E}_{p_{\psi^\star}} (P^m \left(\phi_i - \mathbb{E}_{p_\psi} \phi_i\right))^2\right|}_{\Delta_1} \\ &\quad + 2\underbrace{\left|\mathbb{E}_{p_\psi} \phi_i \, \mathbb{E}_{p_{\psi^\star}} P^m \left(\phi_i - \mathbb{E}_{p_\psi} \phi_i\right)\right|}_{\Delta_2}\end{aligned}$$

where

$$\Delta_1 = \underbrace{\mathbb{E}_{p_\psi}(P_\psi^m(\phi_i - \mathbb{E}_\psi \phi_i))^2}_{\Delta_{1,1}} + \underbrace{\mathbb{E}_{p_\psi}\left(\left(P_\psi^m \phi_i - \mathbb{E}_{p_\psi}\phi_i\right)^2 \left(\frac{\mathrm{d}p_{\psi^\star}}{\mathrm{d}p_\psi} - 1\right)\right)}_{\Delta_{1,2}}$$

$$\overset{(a)}{\leq} \alpha^{2m}\mathbb{E}_\psi(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2 + C_\chi \left\|\psi - \psi^\star\right\| \left(\mathbb{E}_{p_\psi}(P^m(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^4\right)^{1/2}$$

$$\overset{(b)}{\leq} \alpha^{2m}\mathbb{E}_\psi(\phi_i - \mathbb{E}_\psi\phi_i)^2 + C_\chi \left\|\psi - \psi^\star\right\| \left(\mathbb{E}_{p_\psi}P^m(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2\right)^{1/4} \left(\mathbb{E}_{p_\psi}P^m(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^6\right)^{1/4}$$

$$\overset{(c)}{\leq} \alpha^{2m}\mathbb{E}_\psi(\phi_i - \mathbb{E}_\psi\phi_i)^2 + \alpha^{m/2}C_\chi \left\|\psi - \psi^\star\right\| \left(\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2\right)^{1/4} \left(\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^6\right)^{1/4}.$$

In (a), we used the restricted spectral gap Assumption A3 for $\Delta_{1,1}$, and the Cauchy-Schwarz inequality combined with Assumption A2 for $\Delta_{1,2}$. In (b), we used the Cauchy-Schwarz once again, and in (c) we used the fact that P_ψ^m is a contraction in $L^6(p_\psi)$ and another invocation of the spectral gap assumption A3 As an aside, note that a simpler result can be obtained by making regularity assumption on the mapping $\psi \longmapsto P_\psi^m$. Assuming that $\psi \longmapsto P_\psi^m(x)$ is uniformly L_m-Lipschitz across $x \in \mathcal{X}$ for instance (as done in [21, Assumption 5]), the second term $\Delta_{1,2}$ of Δ_1 could have been handled using

$$\begin{aligned}
\Delta_{1,2} &\leq 2\mathbb{E}_{p_{\psi^\star}} \left\|P_\psi^m \phi - \mathbb{E}_{\psi^\star}\phi\right\|^2 + 2\left\|\mathbb{E}_{p_\psi}\phi - \mathbb{E}_{p_{\psi^\star}}\phi\right\|^2 \\
&\leq 4(L_m \left\|\psi - \psi^\star\right\| + \sigma_\star^2\alpha^{2m} + 2\left\|\mathbb{E}_{p_\psi}\phi - \mathbb{E}_{p_{\psi^\star}}\phi\right\|^2) \\
&\leq 4(L_m \left\|\psi - \psi^\star\right\| + \sigma_\star^2\alpha^{2m} + 2L^2\left\|\psi - \psi^\star\right\|)
\end{aligned}$$

Although this result does not require possibly large constants related to sixth-order moments, it is less tight in the sense that it does not go to 0 as $m \to \infty$. Back to the main proof, and Δ_2 in particular. Applying Lemma C.7 to $f := \phi$, we have

$$\begin{aligned}
\Delta_2 &\leq \alpha^m \mathbb{E}_{p_\psi}\phi_i \, C_\chi \left\|\psi - \psi^\star\right\| \left(\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2\right)^{1/2} \\
&\leq \alpha^m C_\chi \partial_i^1 \log Z(\psi)\partial_i^2 \log Z(\psi)^{1/2}\left\|\psi - \psi^\star\right\|
\end{aligned}$$

Putting everything together, we have:

$$\begin{aligned}
|\mathbb{E}_{p_{\psi^\star}}(P_\psi^m\phi_i)^2 - \mathbb{E}_{p_\psi}\phi_i^2| \; &\leq \alpha^{2m}\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2 \\
&+ C_\chi \left\|\psi - \psi^\star\right\| \Big(\alpha^{m/2}(\mathbb{E}_{p_\psi}(\phi_i - \log\phi_i)^6)^{1/4}(\mathbb{E}_{p_\psi}(\phi_i - \log\phi_i)^2)^{1/4} \\
&\qquad + 2\alpha^m \partial_i^1 \log Z(\psi)\partial_i^2 \log Z(\psi)^{1/2}\Big) \\
&\leq \alpha^{2m}\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2 \\
&+ C_\chi\alpha^{m/2}\left\|\psi - \psi^\star\right\| \Big((\mathbb{E}_{p_\psi}(\phi_i - \log\phi_i)^6)^{1/4}(\mathbb{E}_{p_\psi}(\phi_i - \log\phi_i)^2)^{1/4} \\
&\qquad + 2\left|\partial_i^1 \log Z(\psi)\partial_i^2 \log Z(\psi)^{1/2}\right|\Big)
\end{aligned}$$

Summing over i, we obtain

$$\begin{aligned}
&|\mathbb{E}_{p_{\psi^\star}}\left\|P_\psi^m\phi\right\|^2 - (\mathbb{E}_{p_\psi}\phi)^2| \\
&\quad \leq \textstyle\sum_{i=1}^p |\mathbb{E}_{p_{\psi^\star}}(P_\psi^m\phi_i)^2 - (\mathbb{E}_{p_\psi}\phi_i)^2| \\
&\quad \leq \alpha^{2m}\textstyle\sum_{i=1}^p \mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2 + C_\chi\alpha^{m/2}\left\|\log Z\right\|_{2,\infty}\left\|\psi - \psi^\star\right\| \\
&\quad \leq \alpha^{2m}\sigma_\psi^2 + C_\chi\alpha^{m/2}\left\|\log Z\right\|_{2,\infty}\left\|\psi - \psi^\star\right\|
\end{aligned}$$

where

$$\begin{aligned}
\left\|\log Z\right\|_{2,\infty} = \sup_{\psi\in\Psi}\textstyle\sum_{i=1}^p &\left(\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^6\right)^{1/4}\left(\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^2\right)^{1/4} \\
&+ 2\left|\partial_i^1 \log Z(\psi)\partial_i^2 \log Z(\psi)^{1/2}\right|
\end{aligned}$$

Similarly to the previous lemma, one can upper bound $\mathbb{E}(\phi_i - \mathbb{E}_{p_\psi}\phi_i)^6$ using the *centered* moment to cumulant formula:

$$\begin{aligned}\mathbb{E}_{p_\psi}(\phi_i - \mathbb{E}\phi_i)^6 =& 15\partial_i^2 \log Z(\psi)^3 + 10\partial_i^3 \log Z(\psi)^2 + 15\partial_i^2 \log Z(\psi)\partial_i^4 \log Z(\psi) + \partial_i^6 \log Z(\psi) \\ =:& F(\psi)\end{aligned}$$

To get a full description of $\|\log Z\|_{2,\infty}$:

$$\|\log Z\|_{2,\infty} = \sup_{\psi\in\Psi} \sum_{i=1}^p \left(F(\psi)\partial_i^2 \log Z(\psi)\right)^{1/4} + 2\left|\partial_i^1 \log Z(\psi)\partial_i^2 \log Z(\psi)^{1/2}\right| .$$

□

We can now use the previous lemmas to obtain an expression on the second moment of the contrastive divergence gradient estimator, relating it to the one of the stochastic log-likelihood gradient estimator.

Lemma D.3. *Under A1, A2 and A3, we have:*

$$\mathbb{E}\left\|h_t(\psi, X_t)\right\|^2 \leq 2\sigma_\star^2 + 2\sigma_\psi^2 + 2L^2\left\|\psi - \psi^\star\right\|^2 + 4(\sigma_\psi^2\alpha^{2m} + \alpha^{m/2}\left\|\log Z\right\|_{3,\infty} C_\chi \left\|\psi - \psi^\star\right\|)$$

where $\|\log Z\|_{3,\infty} := 2\max(\|\log Z\|_{1,\infty}, \|\log Z\|_{2,\infty})$.

Proof. We rely on the following decomposition:

$$\begin{aligned}h_t(\psi, X_t) \;=\; & \underbrace{(\phi(X_t) - \mathbb{E}_{p_{\psi^\star}}\phi)}_{\Delta_{1,1}} + \underbrace{(\mathbb{E}_{p_{\psi^\star}}\phi - \mathbb{E}_{p_\psi}\phi)}_{\Delta_{1,2}} + \underbrace{(\mathbb{E}_{k_\psi^m(X_t,\cdot)}\phi - \phi(k_\psi^m(X_t,\cdot)))}_{\Delta_2} \\ & + \underbrace{(\mathbb{E}_{p_\psi}\phi - \mathbb{E}_{k_\psi^m(X_t,\cdot)}\phi)}_{\Delta_3}\end{aligned}$$

$\Delta_{1,1} + \Delta_{1,2}$ form the differentiable stochastic gradient g_t of Equation 6. Note that $\Delta_{1,1}$ is mean-zero, and Δ_2 is mean-zero conditionally on X_t. Consequently, $\mathbb{E}\langle\Delta_2, \Delta_3\rangle = \mathbb{E}\langle\Delta_2, \Delta_{1,1}\rangle = \mathbb{E}\langle\Delta_{1,1}, \Delta_{1,2}\rangle = \mathbb{E}\langle\Delta_{1,1}, \Delta_2\rangle = 0$, and the only mixed-terms that remain to be controlled are $\mathbb{E}\langle\Delta_{1,1}, \Delta_3\rangle$ and $\mathbb{E}\langle\Delta_{1,2}, \Delta_3\rangle$. We first control the unmixed terms, and the simple ones first: we have $\mathbb{E}\left\|\Delta_{1,1}\right\|^2 = \sigma_\star^2$, as well as $\mathbb{E}\left\|\Delta_{1,2}\right\|^2 \leq L^2\left\|\psi - \psi^\star\right\|^2$. For Δ_2, we have:

$$\begin{aligned}\mathbb{E}_{k_\psi^m(x,\cdot)}\left\|\Delta_2\right\|^2 &= \mathbb{E}_{k_\psi^m(x,\cdot)}\|\phi(k_\psi^m(x,\cdot))\|^2 - \|\mathbb{E}_{k_\psi^m(x,\cdot)}\phi(k_\psi^m(x,\cdot))\|^2 \\ &= P_\psi^m\|\phi(x)\|^2 - \|P_\psi^m\phi(x)\|^2\end{aligned}$$

We can invoke Lemmas D.2 and D.1, which guarantee

$$\begin{aligned}\left|\mathbb{E}_{p_{\psi^\star}}\|P_\psi^m\phi\|^2 - \left\|\mathbb{E}_{p_\psi}\phi\right\|^2\right| &\leq \alpha^{2m}\sigma_\psi^2 + C_\chi\alpha^{m/2}\left\|\log Z\right\|_{2,\infty}\left\|\psi - \psi^\star\right\| \\ \left|\mathbb{E}_{p_{\psi^\star}}P_\psi^m\|\phi\|^2 - \mathbb{E}_{p_\psi}\|\phi\|^2\right| &\leq \alpha^m C_\chi\left\|\log Z\right\|_{1,\infty}\left\|\psi - \psi^\star\right\|\end{aligned}$$

to obtain

$$\begin{aligned}&\mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k_\psi^m(x,\cdot)}\left\|\Delta_2\right\|^2 \\ &= \mathbb{E}_{p_\psi}\left\|\phi\right\|^2 - \left\|\mathbb{E}\phi\right\|^2 + \alpha^{2m}\sigma_\psi^2 + C_\chi\alpha^{m/2}\left(\left\|\log Z\right\|_{1,\infty} + \left\|\log Z\right\|_{2,\infty}\right)\left\|\psi - \psi^\star\right\| \\ &= \mathbb{E}\left\|\phi - \mathbb{E}\phi\right\|^2 + \alpha^{2m}\sigma_\psi^2 + \alpha^{m/2}\left\|\log Z\right\|_{3,\infty} C_\chi\left\|\psi - \psi^\star\right\|\end{aligned}$$

where $\|\log Z\|_{3,\infty} := 2\max(\|\log Z\|_{1,\infty}, \|\log Z\|_{2,\infty})$.

For Δ_3, notice that Δ_3 is precisely the term Δ_1 in Lemma D.2, and we can thus bound it by

$$\mathbb{E}_{p_{\psi^\star}}\left\|\Delta_3\right\|^2 \leq \sigma_\psi^2\alpha^{2m} + \alpha^{m/2}\left\|\log Z\right\|_{2,\infty} C_\chi\left\|\psi - \psi^\star\right\|$$

Finally, we simply bound $2\mathbb{E}\langle\Delta_{1,1}, \Delta_3\rangle$ by $\mathbb{E}\left\|\Delta_{1,1}\right\|^2 + \mathbb{E}\left\|\Delta_3\right\|^2$, and $2\mathbb{E}\langle\Delta_{1,2}, \Delta_3\rangle$ by $\mathbb{E}\left\|\Delta_{1,2}\right\|^2 + \mathbb{E}\left\|\Delta_3\right\|^2$. Putting everything together, we have:

$$\begin{aligned}\mathbb{E}\left\|h_t(\psi)\right\|^2 &= \mathbb{E}\left\|\Delta_{1,1}\right\|^2 + \mathbb{E}\left\|\Delta_{1,2}\right\|^2 + \mathbb{E}\left\|\Delta_2\right\|^2 + \mathbb{E}\left\|\Delta_3\right\|^2 + 2\mathbb{E}\langle\Delta_{1,1}, \Delta_3\rangle + 2\mathbb{E}\langle\Delta_{1,2}, \Delta_3\rangle \\ &\leq 2\mathbb{E}\left\|\Delta_{1,1}\right\|^2 + 2\mathbb{E}\left\|\Delta_{1,2}\right\|^2 + \mathbb{E}\left\|\Delta_2\right\|^2 + 3\mathbb{E}\left\|\Delta_3\right\|^2 \\ &\leq 2\sigma_\star^2 + 2L^2\left\|\psi - \psi^\star\right\|^2 + \sigma_\psi^2 + 4(\sigma_\psi^2\alpha^{2m} + \alpha^{m/2}\left\|\log Z\right\|_{3,\infty} C_\chi\left\|\psi - \psi^\star\right\|) \\ &\leq 2\sigma_\star^2 + 2\sigma_\psi^2 + 2L^2\left\|\psi - \psi^\star\right\|^2 + 4(\sigma_\psi^2\alpha^{2m} + \alpha^{m/2}\left\|\log Z\right\|_{3,\infty} C_\chi\left\|\psi - \psi^\star\right\|)\end{aligned}$$

□

D.2 Proof of the SGD recursion (Lemma 3.1)

We are now ready to provide an SGD-style recursion for the expected squared distance to the optimum $\mathbb{E}\left[\|\psi_t - \psi^\star\|^2\right]$.

Lemma (Restatement of Lemma 3.1)**.** *Let ψ_t be the iterates produced by Algorithm 1. Denote $\delta_t = \mathbb{E}\left[\|\psi_t - \psi^\star\|^2\right]$, $\sigma_\star = (\mathbb{E}_{p_{\psi^\star}}\|\phi - \mathbb{E}_{p_{\psi^\star}}\phi\|^2)^{1/2}$, and $\sigma_t = (\mathbb{E}_{p_{\psi_t}}\|\phi - \mathbb{E}_{p_{\psi_t}}\phi\|^2)^{1/2}$. Then, under Assumptions A1, A2 and A3, for all $t \geq 1$, we have:*

$$\delta_t \leq (1 - 2\eta_t\tilde{\mu}_{m,t-1} + 2\eta_t^2 L^2)\delta_{t-1} + \eta_t^2\tilde{\sigma}_{m,t-1}^2 + 4\alpha^{m/2}\eta_t^2 \|\log Z\|_{2,\infty} C_\chi \delta_{t-1}^{1/2}$$

where $\|\log Z\|_{3,\infty}$ is a constant, $\tilde{\mu}_{m,t} \coloneqq \mu - \alpha^m\sigma_t C_\chi$, and $\tilde{\sigma}_{m,t} \coloneqq (\sigma_\star^2 + \sigma_t^2 + 2\sigma_t^2\alpha^{2m})^{1/2}$.

Proof. In this proof, we note $(\mathcal{F}_t)_{t\geq 0}$, the increasing family of σ-algebras generated by the random variables $(X_t)_{t\geq 0} \sim p_{\psi^\star}$ and the Markov chain samples $\tilde{X}_t^m | X_t, \psi_t \sim k_{\psi_t}^m(X_t, \cdot)$. We decompose the integrand of δ_t as follows:

$$\begin{aligned}
\|\psi_t - \psi_n^\star\|^2 &= \|\mathrm{Proj}_\Psi(\psi_{t-1} - \eta_t h_t(\psi_{t-1})) - \psi_n^\star\|^2 \\
&\leq \|\psi_{t-1} - \eta_t h_t(\psi_{t-1}) - \psi_n^\star\|^2 \quad \text{(By Lemma C.4)} \\
&= \|\psi_{t-1} - \psi^\star\|^2 - 2\eta_t \langle h_t(\psi_{t-1}), \psi_{t-1} - \psi_n^\star\rangle + \eta_t^2 \|h_t(\psi_{t-1})\|^2
\end{aligned}$$

The first term is (up to an averaging operation) the previous iterate. The middle term will ensure (provided m is large enough) contraction of the expected distance to the optimum. Finally, the third term can be described by Lemma D.3, and essentially behaves like the second moment of a log-likelihood stochastic gradient. Indeed, noting $g(\psi) \coloneqq -\mathbb{E}_{p_{\psi^\star}}\phi + \mathbb{E}_{p_\psi}\phi$ the expectation of g_t w.r.t x_t (which is the gradient of the negative cross-entropy between p_ψ and $p_{\psi^\star}$), we have:

$$\langle h_t(\psi_{n,t-1}), \psi_{t-1} - \psi_n^\star\rangle = \langle g(\psi_{n,t-1}), \psi_{t-1} - \psi_n^\star\rangle + \underbrace{\langle (h_t(\psi_{n,t-1}) - g(\psi_{n,t-1})), \psi_{t-1} - \psi_n^\star\rangle}_{\Delta}$$

and applying Lemma C.7, we get that

$$\begin{aligned}
h_t(\psi_{t-1}) - g(\psi_{t-1}) &= \phi(k^m(X_t, \cdot)) - \mathbb{E}_{p_\psi}\phi \\
\implies \mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k_\psi^m(X_t,\cdot)}\left[h_t(\psi_{t-1}) - g(\psi_{t-1}) | \mathcal{F}_{t-1}\right] &= P_\psi^m\phi - \mathbb{E}_{p_\psi}\phi\ ,
\end{aligned}$$

meaning

$$\begin{aligned}
|\mathbb{E}\left[\Delta|\mathcal{F}_{t-1}\right]| &\leq \left\|\mathbb{E}_{p_{\psi^\star}} P_\psi^m(\phi - \mathbb{E}_{p_\psi}\phi)\right\| \|\psi_{t-1} - \psi^\star\| \\
&\leq \alpha^m C_\chi (\mathbb{E}_{p_{\psi_{t-1}}} \left\|f - \mathbb{E}_{p_{\psi_{t-1}}} f\right\|^2)^{1/2} \|\psi - \psi^\star\|^2 \\
&\leq \alpha^m \sigma_{t-1} C_\chi \|\psi - \psi^\star\|^2
\end{aligned}$$

On the other hand, by applying Lemma C.8 to g, we have:

$$\langle g(\psi_{t-1}), \psi_{t-1} - \psi^\star\rangle = \langle g(\psi_{t-1}) - g(\psi^\star), \psi_{t-1} - \psi^\star\rangle \geq \mu \|\psi_{t-1} - \psi^\star\|^2$$

Combining the above results, we obtain:

$$\begin{aligned}
&\mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k_\psi^m(x,\cdot)}\left[\|\psi_t - \psi^\star\|^2 |\mathcal{F}_{t-1}\right] \\
&\leq (1 - 2\eta_t(\mu - \alpha^m\sigma_{t-1}C_\chi))\|\psi_{t-1} - \psi^\star\|^2 \\
&\quad + \eta_t^2(2\sigma_\star^2 + 2\sigma_{t-1}^2 + 2L^2\|\psi_{t-1} - \psi^\star\|^2 + 4(\sigma_{t-1}^2\alpha^{2m} + \alpha^{m/2}\|\log Z\|_{3,\infty} C_\chi \|\psi_{t-1} - \psi^\star\|))
\end{aligned}$$

And the result follows by integrating over $\mathcal{F}_{t-1}$. □

D.3 Proof of Online CD convergence

We now prove Theorem 3.2. The recursion of Lemma 3.1 is almost identifiable, up to a cross-term of second order, with the one of an SGD algorithm as presented in the setting of [22, Theorem 1].

To make the identification exact, we use the bound $4\alpha^{m/2}\eta_t^2 \|\log Z\|_{3,\infty} C_\chi \delta_{t-1}^{1/2} \leq 2\alpha^{m/2}\eta_t^2\delta_t + 2\alpha^{m/2}\eta_t^2 \|\log Z\|_{3,\infty}^2 C_\chi^2$, yielding the following recursion:

$\delta_t \leq (1 - 2\eta_t(\mu - \alpha^m\sigma C_\chi) + 2\eta_t^2(L^2+\alpha^{m/2}))\delta_{t-1} + (\sigma^2(2+2\alpha^{2m})+\alpha^{m/2}\|\log Z\|_{3,\infty}^2 C_\chi^2)\eta_t^2$

where we used the fact that $\tilde{\sigma}_{m,t} \leq \sigma$. This recursion is of the same form as the one studied in [22, Equation 6, Theorem 1] given by:

$$\delta_t \leqslant \left(1 - 2\mu\gamma_t + 2L^2\gamma_t^2\right)\delta_{t-1} + 2\sigma^2\gamma_t^2$$

by identifying:

$$\begin{aligned}
\sigma^2 &\leftarrow \sigma^2(2+2\alpha^{2m}) + \alpha^{m/2}\|\log Z\|_{3,\infty}^2 =: \tilde{\sigma}_m^2 \\
L^2 &\leftarrow (L^2 + \alpha^{m/2}) =: \tilde{L}^2 \\
\mu &\leftarrow \mu - \alpha^m\sigma C_\chi =: \tilde{\mu}_m \\
\gamma_t &\leftarrow \eta_t
\end{aligned}$$

We can use the same unrolling strategy as theirs (the only condition required to proceed is that $\tilde{\mu}_m > \tilde{L}$, which automatically holds since $\mu < L$), and we obtain

$$\delta_n \leqslant \begin{cases} 2\exp\left(4\tilde{L}C^2\varphi_{1-2\beta}(n)\right)\exp\left(-\frac{\tilde{\mu}_m C}{4}n^{1-\beta}\right)\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + \frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m n^\beta}, & \text{if } 0 \leqslant \beta < 1 \\ \frac{\exp\left(2\tilde{L}^2C^2\right)}{n^{\tilde{\mu}_m C}}\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + 2\tilde{\sigma}_m^2 C^2 \frac{\varphi_{\tilde{\mu}_m C/2-1}(n)}{n^{\tilde{\mu}_m C/2}}, & \text{if } \beta = 1. \end{cases}$$

□

D.4 Proof of online CD with averaging (Theorem 3.3)

We first restate the theorem in its complete form.

Theorem (Contrastive Divergence with Polyak-Ruppert averaging)**.** *Let $(\psi_t)_{t\geq 0}$ the sequence of iterates obtained by running the CD algorithm with a learning rate $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$. Define $\bar{\psi}_n := \frac{1}{n}\sum_{i=1}^n \psi_i$. Then, under the same assumptions as Theorem 3.2 we have, for all $n \geq 1$,*

$$\sqrt{\mathbb{E}\left\|\overline{\psi}_n - \psi^\star\right\|^2} \;\leq\; 2\sqrt{\frac{\mathrm{tr}(\mathcal{I}(\psi)^{-1})}{n}} + \mathcal{O}\left(n^{\max\left(-\left(\frac{1}{2}+\frac{\beta}{4}\right), -\beta, \frac{\beta}{2}-1, -\left(\frac{\beta}{2}+m\frac{|\log\alpha|}{\log n}\right)\right)}\right)$$

Where $\mathcal{I}(\psi^\star) := \mathrm{Cov}_{p_{\psi^\star}}[\phi]$ is the Fisher information matrix of the data distribution. Additionally, if $m > \frac{(1-\beta)\log n}{2|\log\alpha|}$, we have $\sqrt{\mathbb{E}\left\|\overline{\psi}_n - \psi^\star\right\|^2} \;\leq\; 2\sqrt{\frac{\mathrm{tr}(\mathcal{I}(\psi)^{-1})}{n}} + o\left(n^{-1/2}\right)$.

Throughout the proof, we will denote by h_n the standard online CD gradient defined in Equation 3:

$$h_n(\psi_{n-1}) = -\phi(x_n) + \phi(k^m_{\psi_{n-1}}(\cdot, x_n)), \quad x_n \sim p_{\psi^\star}, \quad \forall n \in \mathbb{N}\setminus\{0\},$$

as well as

$$\overline{h}(\psi_{n-1}) := \mathbb{E}\left[h(\psi_{n-1}) \,|\, \psi_{n-1}\right] = \mathbb{E}_{p_{\psi^\star}}\phi(x_n) + \mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k^m_{\psi_{n-1}}}\phi(k^m_{\psi_{n-1}}(x_n, \cdot)).$$

We start by establishing some intermediate lemmas.

Lemma D.4. *Under Assumptions A1, A2, A3, the online CD iterates produced by Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify*

$$\frac{1}{n}\sqrt{\sum_{i=1}^n \left(\mathbb{E}\left[\|\psi_i - \psi^\star\|^2\right]\right)^{1/2}} = \mathcal{O}(n^{-\frac{1}{2}-\frac{\beta}{4}}).$$

Proof. Let us note $\delta_n = \mathbb{E}\left[\|\psi_n - \psi^\star\|^2\right]$. Applying $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ to the r.h.s of Theorem 3.2, we have

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n \delta_i^{1/2} &\leq \frac{2C^{1/2}\tilde{\sigma}_m}{2\tilde{\mu}_m^{1/2}n^2}\sum_{i=1}^n i^{-\beta/2} + \frac{\sqrt{2}}{n^2}\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right)\underbrace{\sum_{i=1}^n e^{2\tilde{L}^2C^2\varphi_{1-2\beta}(i)}e^{-\frac{\tilde{\mu}_m C}{8}i^{1-\beta}}}_{A_3} \\
&\leq \frac{2C^{1/2}\tilde{\sigma}_m}{n^2\tilde{\mu}_m^{1/2}}\varphi_{1-\beta/2}(n) + \frac{\sqrt{2}}{n^2}\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{L^2}\right)A_3 \\
\implies \frac{1}{n}\sqrt{\sum_{i=1}^n \delta_i^{1/2}} &\leq \frac{1}{n}\sqrt{\frac{2C^{1/2}\tilde{\sigma}_m}{\tilde{\mu}_m^{1/2}}\varphi_{1-\beta/2}(n)} + \frac{1}{n}\left(\sqrt{2}(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2})A_3\right)^{1/2} = \mathcal{O}\left(n^{-\frac{1}{2}-\frac{\beta}{4}}\right).
\end{aligned}$$

where A_3 is finite if $\beta < 1$, and $A_3 = O(n)$ otherwise [22]. $\square$

Lemma D.5. *Under Assumptions A1, A2, A3, the online CD iterates produced by Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify*

$$\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n h_i(\psi_{i-1})\right\|^2\right]} = \mathcal{O}\left(n^{\frac{\beta}{2}-1}\right).$$

Proof. The result follows from the fact that h_t verifies

$$h_t(\psi_{t-1}) \;=\; \frac{1}{\eta_t}(\psi_{t-1} - \psi_t)\;.$$

A similar quantity was handled in the case of standard SGD [22, Theorem 3], and the only condition needed to reuse their steps is that $(\psi_t)_{t\leq n}$ satisfies an upper bound of the same form as the one [22, Theorem 1] derived. This is precisely the nature of our bound of ψ_t estblished in Theorem 3.2, with $\tilde{\mu}_m, \tilde{\sigma}_m, \tilde{L}$. Borrowing on their result, we have:

$$\begin{aligned}\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n h_i(\psi_{i-1})\right\|^2\right]} \leq\; & \frac{4\tilde{\sigma}_m\beta}{C^{1/2}n\tilde{\mu}_m}\varphi_{\beta/2}(n) + \frac{4\beta}{Cn\tilde{\mu}_m^{1/2}}\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right)^{1/2}A_2 \\ & + \frac{1}{n\tilde{\mu}_m^{1/2}}\left(\frac{1}{C} + 2\tilde{L}\right)\delta_0^{1/2} + \frac{2\tilde{L}}{n\tilde{\mu}_m^{1/2}}\frac{2C^{1/2}\tilde{\sigma}_m}{\tilde{\mu}_m^{1/2}}\varphi_{1-\beta}(n)^{1/2} \\ & + \frac{4\tilde{L}}{n\tilde{\mu}_m^{1/2}}\left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right)^{1/2}A_2^{1/2}\end{aligned}$$

where $\tilde{\mu}_m, \tilde{\sigma}_m$ and $\tilde{L}$ are defined in 3.2, and

$$A_2 \;=\; \sum_{k=1}^n e^{\frac{-\tilde{\mu}_m C}{16}k^{1-\beta} + 16\tilde{L}_1^4 C^4 \varphi_{1-2\beta}(k)}$$ $\square$

Lemma D.6. *Under Assumptions A1, A2, A3, the online CD iterates produced by Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify*

$$\frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}\left[\|\psi_i - \psi^\star\|^4\right]} = \mathcal{O}(n^{-\beta}).$$

Proof. We proceed as in the proof of [22, Theorem 3], first establishing a recurrence for $\mathbb{E}\left[\|\psi_i - \psi^\star\|^4\right]$, and then unrolling it. We have

$$\begin{aligned}\mathbb{E}\left[\|\psi_n - \psi^*\|^4 \mid \mathcal{F}_{n-1}\right] \leqslant\; & \|\psi_{n-1} - \psi^*\|^4 + 6\eta_n^2\|\psi_{n-1} - \psi^*\|^2\,\mathbb{E}\left[\|h_n(\psi_{n-1})\|^2 \mid \mathcal{F}_{n-1}\right] \\ & + \eta_n^4\mathbb{E}\left[\|h_n(\psi_{n-1})\|^4 \mid \mathcal{F}_{n-1}\right] \\ & - 4\eta_n\|\psi_{n-1} - \psi^*\|^2\langle\psi_{n-1} - \psi^*, \mathbb{E}[h_n]\rangle \\ & + 4\eta_n^3\|\psi_{n-1} - \psi^*\|\,\mathbb{E}\left[\|h_n(\psi_{n-1})\|^3 \mid \mathcal{F}_{n-1}\right].\end{aligned}$$

The second and fourth terms will be controlled using results from our previous sections. For simplicity, we don't attempt to relate the moments of $\|h_n\|^4$ as precisely as before. Instead we use:

$$\mathbb{E}_{p_{\psi^\star}}\left[\|h_n\|^k\right] \leq 2^{k-1}\left(\mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k^m_{p_{\psi_{n-1}}}}\left[\|\phi(k^m_{\psi_{n-1}}(x_n,\cdot))\|^k\right] + \mathbb{E}_{p_{\psi^\star}}\left[\|\phi(x_n)\|^k\right]\right)$$

And let us note $\tau = \left(\sup_{\psi\in\Psi}\mathbb{E}\|\phi\|^8\right)^{1/8}$. We have, for $k \geq 4$,

$$\begin{aligned}\mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k^m_{p_{\psi_{n-1}}}}\|\phi(k^m_{\psi_{n-1}}(x_n,\cdot))\|^k &\leq \mathbb{E}_{p_{\psi_{n-1}}}\|\phi\|^k + C_\chi(\mathbb{E}(\|\phi\|^k - \mathbb{E}\|\phi\|^k)^2)^{1/2}\|\psi_{n-1} - \psi^\star\| \\ &\leq \tau^k + 2C_\chi\tau^k\|\psi_{n-1} - \psi^\star\|\end{aligned}$$

Where we used the fact that $P^m_{\psi_{n-1}}$ is a contraction, and $(\mathbb{E}\|\phi\|^k)^{1/k}$ is an increasing function of k. On the other hand, we simply have $\mathbb{E}_{p_{\psi^\star}}\|\phi(x_n,\cdot)\|^k \leq \tau^k$. Plugging this into the previous equation, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\psi_n-\psi^*\|^4 \mid \mathcal{F}_{n-1}\right] \leqslant & \|\psi_{n-1}-\psi^*\|^4 \\ & + 6\eta_n^2\|\psi_{n-1}-\psi^*\|^2\left(4\tau^2+4C_\chi\tau^2\|\psi_{n-1}-\psi^\star\|\right) \\ & + \eta_n^4(16\tau^4+16C_\chi\tau^4\|\psi_{n-1}-\psi^\star\|) \\ & - 4\eta_n\|\psi_{n-1}-\psi^*\|^2\langle\psi_{n-1}-\psi^*,\mathbb{E}[h_n]\rangle \\ & + 4\eta_n^3\|\psi_{n-1}-\psi^*\|\left(8\tau^3+8C_\chi\tau^3\|\psi_{n-1}-\psi^\star\|\right)\end{aligned}$$

To simplify the recursion, we use the four following inequalities:

$$\begin{aligned}\tau^2\eta_n^2\|\psi_{n-1}-\psi^\star\|^3 &\leq \frac{1}{2}(\tau^2\eta_n^2(\|\psi_{n-1}-\psi^\star\|^2+\tau^2\eta_n^2\|\psi_{n-1}-\psi^\star\|^4) \\ \tau^4\eta_n^4\|\psi_{n-1}-\psi^\star\| &\leq (\tau^4\eta_n^4+\frac{1}{4}\tau^4\eta_n^4\|\psi_{n-1}-\psi^\star\|^4)) \\ \eta_n^3\tau^3\|\psi_{n-1}-\psi^\star\| &\leq \frac{1}{2}(\eta_n^2\tau^2\|\psi_{n-1}-\psi^\star\|^2+16\eta_n^4\tau^4 \\ \tau^3\eta_n^3\|\psi_{n-1}-\psi^\star\|^2 &\leq \frac{1}{2}(\eta_n^4\tau^4+\|\psi_{n-1}-\psi^\star\|^2\eta_n^2\tau^2)\end{aligned}$$

Injecting them in our recursion, we obtain:

$$\begin{aligned}\mathbb{E}\left[\|\psi_n-\psi^*\|^4 \mid \mathcal{F}_{n-1}\right] \leqslant & \|\psi_{n-1}-\psi^*\|^4 \\ & + 12\eta_n^2\|\psi_{n-1}-\psi^*\|^2\tau^2 \\ & + 12C_\chi\tau^2\eta_n^2\|\psi_{n-1}-\psi^\star\|^2 \\ & + 12C_\chi\tau^2\eta_n^2\|\psi_{n-1}-\psi^\star\|^4 \\ & + 16\eta_n^4\tau^4 \\ & + 16C_\chi\eta_n^4\tau^4 \\ & + 4C_\chi\eta_n^4\tau^4\|\psi_{n-1}-\psi^\star\|^4 \\ & - 4\eta_n\tilde{\mu}_m\|\psi_{n-1}-\psi^*\|^4 \\ & + 16\eta_n^2\tau^2\|\psi_{n-1}-\psi^*\|^2 \\ & + 16\eta_n^4\tau^4 \\ & + 16\eta_n^4C_\chi\tau^4 \\ & + 16\eta_n^2C_\chi\tau^2\|\psi_{n-1}-\psi^\star\|^2 \ ,\end{aligned}$$

which, after further simplifications, yields

$$
\begin{aligned}
&\mathbb{E}\big[\,\|\psi_n-\psi^*\|^4 \mid \mathcal{F}_{n-1}\big] \\
&\quad\le \|\psi_{n-1}-\psi^\star\|^4\,(1-4\eta_n\tilde\mu_m+12C_\chi\eta_n^2\tau^2+4C_\chi\tau^4\eta_n^4) \\
&\quad+\eta_n^2\,\|\psi_{n-1}-\psi^\star\|^2\,(28(1+C_\chi)\tau^2)+32\eta_n^4(\tau^4(1+C_\chi)) \\
&\quad\le \|\psi_{n-1}-\psi^\star\|^4\,(1-4\eta_n\tilde\mu_m+12(1+C_\chi)\eta_n^2\tau^2+4(1+C_\chi)\tau^4\eta_n^4) \\
&\quad+28\eta_n^2\,\|\psi_{n-1}-\psi^\star\|^2\,((1+C_\chi)\tau)^2+32\eta_n^4(1+C_\chi)\tau)^4 \\
&\quad\le \|\psi_{n-1}-\psi^\star\|^4\,(1-4\eta_n\tilde\mu_m+12\eta_n^2((1+C_\chi)\tau)^2+4((1+C_\chi)\tau)^4\eta_n^4) \\
&\quad+28\eta_n^2\,\|\psi_{n-1}-\psi^\star\|^2\,((1+C_\chi)\tau)^2+32\eta_n^4(1+C_\chi)\tau)^4 \\
&\quad\le \|\psi_{n-1}-\psi^\star\|^4\,(1-4\eta_n\tilde\mu_m+12\eta_n^2(2(1+C_\chi)\tau)^2+16\eta_n^2(2(1+C_\chi)\tau)^3 \\
&\quad+4(2(1+C_\chi)\tau)^4\eta_n^4)+20\eta_n^2\,\|\psi_{n-1}-\psi^\star\|^2\,(2(1+C_\chi)\tau)^2+16\eta_n^4(2(1+C_\chi)\tau)^4 \\
&\quad\le \|\psi_{n-1}-\psi^\star\|^4\,(1-4\eta_n\tilde\mu_m+12\eta_n^2(2(1+C_\chi)\tau+L)^2+16\eta_n^2(2(1+C_\chi)\tau+L)^3 \\
&\quad+4(2(1+C_\chi)\tau+L)^4\eta_n^4)+20\eta_n^2\,\|\psi_{n-1}-\psi^\star\|^2\,(2(1+C_\chi)\tau)^2+16\eta_n^4(2(1+C_\chi)\tau)^4 \\
&\quad\le \|\psi_{n-1}-\psi^\star\|^4\,(1-4\eta_n\tilde\mu_m+12\eta_n^2\tilde L_1^2+16\eta_n^2\tilde L_1^3+4\tilde L_1^4\eta_n^4) \\
&\quad+20\eta_n^2\,\|\psi_{n-1}-\psi^\star\|^2\,\tilde\tau_1^2+16\eta_n^4\tilde\tau_2^4
\end{aligned}
$$

where we defined $\tilde\tau_1 \coloneqq 2(1+C_\chi)\tau$ and $\tilde L_1 \coloneqq 2(1+C_\chi)\tau+L$. This recursion is of the form of the one studied in [22, Equation 32] (note that by design, $\tilde L \ge \tilde\mu_m$.) The steps performed to bound $\mathbb{E}\left[\|\psi_i-\psi^\star\|^4\right]$ thus follow from their derivations, and we obtain:

$$
\begin{aligned}
&\frac{1}{n}\sqrt{\textstyle\sum_{i=1}^n \mathbb{E}\,\|\psi_{i-1}-\psi^\star\|^4} \\
&\leqslant \frac{C\tilde\tau_1^2}{2n}\Big(C^{1/2}\varphi_{1-3\beta/2}(n)+\tilde\mu_m^{-1/2}\varphi_{1-\beta}(n)\Big) \\
&\quad+\frac{\sqrt{20}C^{1/2}\tilde\tau_1}{2n}A_1\exp\Big(24\tilde L_1^4C^4\Big)\Big(\delta_0+\frac{\tilde\mu_m\mathbb{E}\,\|\theta_0-\theta^*\|^4}{20C\tilde\tau_1^2}+2\tilde\tau_1^2C^3\tilde\mu_m+8\tilde\tau_1^2C^2\Big)^{1/2} \\
&=\mathcal{O}\left(n^{-\beta}\right),
\end{aligned}
$$

where

$$A_1=\textstyle\sum_{k=1}^n e^{\frac{-\tilde\mu C}{16}k^{1-\beta}+16\tilde L_1^4C^4\varphi_{1-2\beta}(k)}$$

and we have $A(1)<+\infty$ if $\beta<1$, and $A(1)=O(n)$ otherwise. □

Lemma D.7. *For all $\psi\in\Psi$, we have:*

$$\left\|\mathrm{Cov}\left[\phi(k_\psi^m(x_n,\cdot))\right]-\mathrm{Cov}_{p_\psi}\left[\phi(x_n)\right]\right\|_{\mathrm{F}}\le\alpha^mC_\chi(\bar\sigma+2\|\log Z\|_{3,+\infty}\sigma)\|\psi-\psi^\star\|$$

where $\bar\sigma^2\coloneqq\sup_{\psi\in\Psi}\mathbb{E}_{p_\psi}\left[\left\|\phi\phi^\top-\mathbb{E}_{p_\psi}[\phi\phi^\top]\right\|_F^2\right]<+\infty$ and $\|\log Z\|_{4,\infty}\coloneqq\sup_{\psi\in\Psi}\|\mathbb{E}_\psi\phi\|\le\sup_{\psi\in\Psi}\sum_{i=1}^d\partial_i^2\log Z(\psi)^2$.

Proof. We have

$$\mathrm{Cov}\left[\phi(k_\psi^m(x_n,\cdot))\right]=\mathbb{E}_{p_{\psi^\star}}P_\psi^m\left[\phi\phi^\top\right]-\left(\mathbb{E}_{p_{\psi^\star}}[P_\psi^m\phi]\right)\left(\mathbb{E}_{p_{\psi^\star}}[P_\psi^m\phi]^\top\right)$$

Looking at the second moment first, we have

$$\mathbb{E}_{p_{\psi^\star}}\left[P_\psi^m\phi\phi^\top\right]-\mathbb{E}_{p_\psi}\phi\phi^\top=\underbrace{\mathbb{E}_{p_{\psi^\star}}P_\psi^m(\phi\phi^\top-\mathbb{E}_{p_\psi}\phi\phi^\top)}_{\Delta_1}.$$

Applying lemma C.7 to the function $\phi_i\phi_j-\mathbb{E}_{p_\psi}\phi_i\phi_j$, over $i,j\in[d]\times[d]$, and taking a square root, we obtain

$$\|\Delta_1\|_{\mathrm{F}}\le\|\psi-\psi^\star\|\,\alpha^mC_\chi\sqrt{\textstyle\sum_{i,j=1}^n\mathbb{E}_{p_\psi}\left[\left\|\phi_i\phi_j^\top-\mathbb{E}_{p_\psi}\phi_i\phi_j^\top\right\|_F\right]}\le\bar\sigma\alpha^mC_\chi\,\|\psi-\psi^\star\|\,.$$

We now investigate the first moment. We have

$$\mathbb{E}_{p_{\psi^\star}}\left[P_\psi^m \phi\right] = \underbrace{\mathbb{E}_{p_{\psi^\star}}\left[P_\psi^m \phi\right] - \mathbb{E}_{p_\psi}\left[\phi\right]}_{\Delta_{2,1}} + \mathbb{E}_{p_{\psi_n}} \phi$$
$$\implies \underbrace{(\mathbb{E}_{p_{\psi^\star}} P_\psi^m \phi)(\mathbb{E}_{p_{\psi^\star}} P_\psi^m \phi)^\top - \mathbb{E}_{p_\psi}\phi\mathbb{E}_{p_\psi}\phi^\top}_{\Delta_2} = \Delta_{2,1}\mathbb{E}_{p_\psi}\phi^\top + \mathbb{E}_{p_\psi}\phi\Delta_{2,1}^\top$$

and thus, applying Lemma C.7 on Δ_2, we have:

$$\begin{aligned}\|\Delta_2\|_{\mathrm{F}} &\leq \left\|\Delta_{2,1}\mathbb{E}_{p_\psi}\left[\phi^\top\right] + \mathbb{E}_{p_\psi}\left[\phi\right]\Delta_{2,1}^\top\right\|_{\mathrm{F}} \\ &\leq 2\left\|\Delta_{2,1}\right\| \left\|\mathbb{E}_{p_\psi}\left[\phi\right]\right\| \\ &\leq 2\left\|\log Z\right\|_{4,\infty} \alpha^m C_\chi \sigma \left\|\psi - \psi^\star\right\|\end{aligned}$$

We can now combine our two matrix moment bounds to obtain

$$\left\|\mathrm{Cov}\left[\phi(k_\psi^m(x_n,\cdot))\right] - \mathrm{Cov}_{p_\psi}\left[\phi(x_n)\right]\right\|_{\mathrm{F}} = \|\Delta_1 + \Delta_2\|_{\mathrm{F}} \leq \alpha^m C_\chi(\bar{\sigma} + 2\|\log Z\|_{4,\infty}\sigma)\|\psi - \psi^\star\|.$$

□

Lemma D.8. *Under Assumptions A1, A2, A3 it holds that*

$$\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n f''(\psi^\star)^{-1}(\overline{h}(\psi_{i-1}) - h_i(\psi_{i-1}))\right\|^2\right]} \leq 2\sqrt{\frac{\mathrm{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + \frac{\|\mathcal{I}(\psi^\star)^{-2}\|_{\mathrm{F}}^{1/2}\left(M + \alpha^m C_\chi(\bar{\sigma} + 2\|\log Z\|_{4,\infty}\sigma)\right)^{1/2}}{n}\left(\sum_{i=1}^n \delta_{i-1}^{1/2}\right)^{1/2},$$

where $M \coloneqq \sup_{\psi\in\Psi}\left\|\nabla^3 \log Z(\psi)\right\|_{\mathrm{op}(\|\cdot\|_F,\|\cdot\|_F)} < +\infty$.

Proof.

$$\begin{aligned}&f''(\psi^\star)^{-1}(\overline{h}(\psi_{n-1}) - h_n(\psi_{n-1})) \\ &\quad= f''(\psi^\star)^{-1}\underbrace{(\phi(x_n) - \mathbb{E}_{p_{\psi^\star}}\phi)}_{\Delta_{1,n}} + f''(\psi^\star)^{-1}\underbrace{(\phi(k_{\psi_{n-1}}^m(x_n)) - \mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k_{\psi_{n-1}}^m}\phi(k_{\psi_{n-1}}^m(x_n,\cdot)))}_{\Delta_{2,n}} \\ &\quad= f''(\psi^\star)^{-1}\Delta_{1,n} + f''(\psi^\star)^{-1}\Delta_{2,n}\end{aligned}$$

Noting Δ the l.h.s of Lemma D.8, we have, summing over $[n]$, and using Minkowski's inequality,

$$\Delta \leq \frac{1}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n f''(\psi^\star)^{-1}\Delta_{1,i}\right\|^2} + \frac{1}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n f''(\psi^\star)^{-1}\Delta_{2,i}\right\|^2}$$

Note that this step was made possible because we are looking at the square-root of the variance, which is unlike the recursion in Lemma 3.1. This allows to separate the terms and use fewer intermediaries than in the proof of Lemma 3.1.

Since both $\Delta_{1,n}$ and $\Delta_{2,n}$ are martingale differences with respect to the filtration $\mathcal{F}_{n-1}$, the covariance terms vanish, and we have

$$
\begin{aligned}
\Delta &\leq \frac{1}{n}\sqrt{\textstyle\sum_{i=1}^n \mathbb{E}\left[\|f''(\psi^\star)^{-1}\Delta_{1,i}\|^2\right]} + \frac{1}{n}\sqrt{\textstyle\sum_{i=1}^n \mathbb{E}\left[\|f''(\psi^\star)^{-1}\Delta_{2,i}\|^2\right]} \\
&\leq \frac{1}{n}\sqrt{\textstyle\sum_{i=1}^n \operatorname{tr}(f''(\psi^\star)^{-1}\mathbb{E}\left[\Delta_{1,i}\Delta_{1,i}^\top f''(\psi^\star)^{-1})\right]} \\
&\quad + \frac{1}{n}\sqrt{\textstyle\sum_{i=1}^n \operatorname{tr}(f''(\psi^\star)^{-1}\mathbb{E}\left[\Delta_{2,i}\Delta_{2,i}^\top f''(\psi^\star)^{-1})\right]} \\
&\leq \sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + \frac{1}{n}\sqrt{\textstyle\sum_{i=1}^n \operatorname{tr}(f''(\psi^\star)^{-1}\operatorname{Cov}\left[\phi(k^m_{\psi_{i-1}}(x_i))\right] f''(\psi^\star)^{-1})} \\
&\leq \sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + \frac{1}{n}\Big(\textstyle\sum_{i=1}^n \operatorname{tr}(f''(\psi^\star)^{-1}(\operatorname{Cov}_{p_{\psi^\star}}\phi + (\operatorname{Cov}_{p_{\psi_{i-1}}}\phi - \operatorname{Cov}_{p_{\psi^\star}}\phi) \\
&\quad + (\operatorname{Cov}\phi(k^m_{\psi_{i-1}}(x_i)) - \operatorname{Cov}_{p_{\psi_{i-1}}}\phi))f''(\psi^\star)^{-1})\Big)^{1/2} \\
&\leq 2\sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + \frac{\sqrt{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-2})}}{n}\Big(\textstyle\sum_{i=1}^n \|(\operatorname{Cov}_{p_{\psi_{n-1}}}\phi - \operatorname{Cov}_{p_{\psi^\star}}\phi\|_{\mathrm{F}} \\
&\quad + \|\operatorname{Cov}\phi(k^m_{\psi}(x_i)) - \operatorname{Cov}_{p_\psi}\phi)\|_{\mathrm{F}}\Big)^{1/2} \\
&\overset{(a)}{\leq} 2\sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + \frac{\sqrt{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-2})}}{n}(M + \alpha^m C_\chi(\bar{\sigma} + 2\|\log Z\|_{4,\infty}\sigma)^{1/2})\times \\
&\quad \sqrt{\textstyle\sum_{i=1}^n \|\psi_{i-1} - \psi^\star\|} \\
&\overset{(b)}{\leq} 2\sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} + \frac{\sqrt{\operatorname{tr}(\mathcal{I}(\psi^\star)^{-2})}}{n}(M + \alpha^m C_\chi(\bar{\sigma} + 2\|\log z\|_{4,\infty}\sigma))^{1/2}\sqrt{\textstyle\sum_{i=1}^n \delta_i^{1/2}}.
\end{aligned}
$$

In (a) we used Lemma D.8, the cyclicity of the trace, $\operatorname{tr}(A^\top B) \leq \|AB\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}$, and the fact that since $\operatorname{Cov}_{p_{\psi_{i-1}}}[\phi(x_i)] = \nabla^2_\psi \mathcal{L}(\psi_{i-1})$, by analycity of $\mathcal{L}$, there exists a constant $M \coloneqq \sup_{\psi\in\Psi}\left\|\nabla^3 \log Z(\psi)\right\|_{\operatorname{op}(\|\cdot\|,\|\cdot\|_F)}$ such that

$$
\|\nabla^2_\psi \mathcal{L}(\psi_{i-1}) - \nabla^2_\psi \mathcal{L}(\psi^\star)\|_{\mathrm{F}} \leq M\|\psi_{i-1} - \psi^\star\|.
$$

In (b) we used Jensen's inequality to get $\mathbb{E}\left[\|\psi_{i-1} - \psi^\star\|\right] \leq \sqrt{\mathbb{E}\left[\|\psi_{i-1} - \psi^\star\|^2\right]} = \delta_{i-1}^{1/2}$. □

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. It holds that:

$$
\begin{aligned}
f''(\psi^\star)(\psi_{n-1} - \psi^\star) &= f'(\psi_{n-1}) - f'(\psi^\star) + (f''(\psi^\star)(\psi_{n-1} - \psi^\star) - f'(\psi_{n-1}) + f'(\psi^\star)) \\
&= h_n(\psi_{n-1}) - f'(\psi^\star) + (f''(\psi^\star)(\psi_{n-1} - \psi^\star) - f'(\psi_{n-1}) + f'(\psi^\star)) \\
&\quad + (f'(\psi_{n-1}) - \bar{h}(\psi_{n-1})) + (\bar{h}(\psi_{n-1}) - h_n(\psi_{n-1})).
\end{aligned}
$$

Applying on both sides: (a) a summation over $i \in [n]$, (b) a multiplication by $f''(\psi^\star)^{-1}$, (c) $\sqrt{\mathbb{E}[\|\cdot\|^2]}$, and using Minkowski's inequality on the r.h.s, we obtain

$$\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \psi_i - \psi^\star\right\|^2\right]} \leq \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n f''(\psi^\star)^{-1}h_i(\psi_{i-1})\right\|^2\right]}}_{(i)}$$
$$+ \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n f''(\psi^\star)^{-1}(f''(\psi^\star)(\psi_{i-1}-\psi^\star) - f'(\psi_{i-1})\right\|^2\right]}}_{(ii)}$$
$$+ \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n f''(\psi^\star)^{-1}(f'(\psi_{i-1}) - \bar{h}(\psi_{i-1}))\right\|^2\right]}}_{(iii)}$$
$$+ \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n f''(\psi^\star)^{-1}(\overline{h}(\psi_{i-1}) - h_i(\psi_{i-1}))\right\|^2\right]}}_{(iv)}.$$

(i) and (ii) have direct analogues in the proofs of prior work [22] on the convergence of (unbiased) SGD with Polyak-Ruppert averaging, and will be bounded similarly. (iii) captures the bias of the CD algorithm, while (iv) captures the variance.

Bounding (i) Using Lemma D.5, we have $(i) = \mathcal{O}(n^{\frac{\beta}{2}-1})$.

Bounding (ii) Since $\log Z(\psi)$ is analytic, there exists some constant M' such that

$$\|f''(\psi^\star)(\psi_{i-1}-\psi^\star) - f'(\psi_{i-1})\| \leq M'\|\psi_{i-1}-\psi^\star\|^2.$$

Thus, we have:

$$(ii) \leq \frac{M'}{n}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^n \|\psi_i - \psi^\star\|^2\right)^2\right]} \leq \frac{M'}{n}\sum_{i=1}^n \sqrt{\mathbb{E}\left[\|\psi_i - \psi^\star\|^4\right]} = \mathcal{O}(n^{-\beta})$$

where the second-to-last inequality used Minkowski's inequality, and the last applied Lemma D.6.

Bounding (iii) By Minkowski's inequality, we have:

$$(\mathbb{E}(iii)^2)^{1/2} \;\leq\; \frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}\left[\left\|f'(\psi_{n-1}) - \bar{h}(\psi_{n-1})\right\|^2\right]}$$

Moreover, using lemma C.7, we have:

$$\begin{aligned}\left\|f'(\psi_{n-1}) - \bar{h}(\psi_{n-1})\right\| &= \left\|\mathbb{E}_{p_{\psi_{n-1}}}\phi - \mathbb{E}_{p_{\psi^\star}}\mathbb{E}_{k^m_{\psi_{n-1}}}\phi\right\| \\ &\leq \alpha^m\sqrt{\mathbb{E}_{p_{\psi_{n-1}}}\left\|\phi - \mathbb{E}_{p_{\psi_{n-1}}}\phi\right\|^2}\, C_\chi \left\|\psi_{n-1}-\psi^\star\right\| \\ &\leq \alpha^m \sigma C_\chi \left\|\psi_{n-1}-\psi^\star\right\|.\end{aligned}$$

We thus obtain

$$(iii) \;\leq\; \frac{\alpha^m C_\chi}{n}\sum_{i=1}^n (\mathbb{E}\left\|\psi_{i-1}-\psi^\star\right\|^2)^{1/2} \;=\; \frac{\alpha^m C_\chi}{n}\sum_{i=1}^n \delta_{i-1}^{1/2}.$$

Recalling that δ_i satisfies Theorem 3.2, we have that $\delta_n^{\frac{1}{2}} = \mathcal{O}(n^{-\beta/2})$, and we thus have $\sum_{i=1}^n \delta_i^{1/2} = \mathcal{O}(n^{1-\frac{\beta}{2}})$. Overall, we obtain that $(iii) = \mathcal{O}(\alpha^m n^{-\beta/2}) = \mathcal{O}(n^{-(\frac{\beta}{2}+m\frac{\lfloor\log\alpha\rfloor}{\log n})})$.

Bounding (iv) Combining Lemma D.8 and Lemma D.4, we obtain that

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \psi_i - \psi^\star\right\|^2\right]} \\
&\leq 2\sqrt{\frac{\operatorname{tr}\left(\mathcal{I}(\psi^\star)^{-1}\right)}{n}} + \frac{\left\|\mathcal{I}(\psi^\star)^{-2}\right\|_{\mathrm{F}}^{1/2}\left(M + \alpha^m C_\chi(\bar{\sigma} + 2\|\log Z\|_{4,\infty}\sigma)\right)^{1/2}}{n}\sqrt{\sum_{i=1}^n \delta_{i-1}^{1/2}} \\
&\leq 2\sqrt{\frac{\operatorname{tr}\left(\mathcal{I}(\psi^\star)^{-1}\right)}{n}} + \mathcal{O}(n^{-\frac{1}{2}+\frac{\beta}{4}}).
\end{aligned}$$

Putting everything together, we have that:

$$\sqrt{\mathbb{E}\left\|\overline{\psi}_n - \psi^\star\right\|^2} \leq 2\sqrt{\frac{\operatorname{tr}(\mathcal{I}(\psi)^{-1})}{n}} + \mathcal{O}(n^{\max\left(-\left(\frac{1}{2}+\frac{\beta}{4}\right), -\beta, \frac{\beta}{2}-1, -\left(\frac{\beta}{2}+m\frac{|\log\alpha|}{\log n}\right)\right)}).$$

If, furthermore, $m > \frac{(1-\beta)\log n}{2|\log\alpha|}$, we have

$$\max\left(-\left(\frac{1}{2}+\frac{\beta}{4}\right), -\beta, \frac{\beta}{2}-1, -\left(\frac{\beta}{2}+m\frac{|\log\alpha|}{\log n}\right)\right) < -\frac{1}{2},$$

which concludes the proof. □

E L_2 approximation by auxiliary gradient updates

In this section, we consider different gradient update schemes starting from some random initialization θ_{init}, and control the L_2 distance between the different updates and the deterministic target $\psi^* \in \Psi$.

Notation. Recall the notation that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\psi^*}$, $B = n/N$ and for $\psi \in \Psi$, $K_\psi(x) \sim k_\psi^m(x, \bullet)$. We also write, for $m \in \mathbb{N} \cup \{\infty\}$,

$$K_{1;\psi}^m(x)\,,\ \ldots\ ,\ K_{n;\psi}^m(x) \overset{\text{i.i.d.}}{\sim} k_\psi^m(x, \bullet)$$

Let θ^{init} be some Ψ-valued random initialization that is possibly correlated with $X_1, \ldots, X_n$. We capture the effect of correlation through the following quantities: For $\epsilon > 0$ and $\nu > 2$, let

$$\vartheta_{n,m}^{\text{init}}(\epsilon) \ := \mathbb{P}\left(\frac{\left\|\sum_{i=1}^n \mathbb{E}\left[\phi\left(K_{\theta^{\text{init}}}^m(X_i)\right) \mid X_i, \theta^{\text{init}}\right] - \mathbb{E}\left[\phi\left(K_{\theta_{\text{init}}}^m(X_i')\right) \mid \theta^{\text{init}}\right]\right\|}{n} > \epsilon\right),$$

and $\quad \varepsilon_{n,m;\nu}^{\text{init}}(\epsilon) \ := \sqrt{\epsilon^2 + \kappa_{\nu;m}^2\left(\vartheta_{n,m}^{\text{init}}(\epsilon)\right)^{\frac{2}{\nu-2}}}\ .$

We also consider the i.i.d. samples, drawn independently of $X_1, \ldots, X_n$ and on a given $\psi \in \Psi$, as

$$X_1^\psi, \ldots, X_n^\psi \overset{\text{i.i.d.}}{\sim} p_\psi\ .$$

For notational clarity, we shall use $\theta_{m,B}$ to denote parameters arising from an one-step update, where the subscripts m, B represent performing the one-step update with length-m Markov chains and with batch size B. This is to be distinguished from ψ_t elsewhere in the text, which denotes the parameter from the actual multi-step CD algorithm and the subscript t denotes the t-th CD iterate.

Gradient update schemes. We consider five different updates. Let X_1' be an i.i.d. copy of X_1 drawn independently of all other random variables. The SGD-with-replacement update is given by

$$\theta_{m,B}^{\text{SGDw}} \ := \ F_{m,B}^{\text{SGDw}}(\theta^{\text{init}})\ , \quad \text{where } F_{m,B}^{\text{SGDw}}(\psi) \ := \ \psi - \frac{\eta}{B}\sum_{i\in S^w}\left(\phi(X_i) - \phi\left(K_{i;\psi}^m(X_i)\right)\right)$$

and S^w is a uniformly drawn size-B subset of $[n]$. The SGD-without-replacement update, after renormalizing the learning rate, is given by the N-fold function composition

$$\begin{aligned}
\theta_{m,B}^{\text{SGDo}} \ &:= \ F_{m,B;N}^{\text{SGDo}} \circ \ldots \circ F_{m,B;1}^{\text{SGDo}}(\theta^{\text{init}})\ , \\
&\text{where} \quad F_{m,B;j}^{\text{SGDo}}(\psi) \ := \ \psi - \frac{\eta}{NB}\sum_{i\in S_j^o}\left(\phi(X_i) - \phi\left(K_{i;\psi}^m(X_i)\right)\right) \quad \text{for each } j \in [N]\ ,
\end{aligned}$$

and S_j^o's are disjoint size-B random subsets of $[n]$, defined by $(S_1^o, \ldots, S_N^o) = \pi([n])$ for a uniformly drawn element π of the permutation group on n objects. The full-batch gradient update is given by

$$\theta_m^{\mathrm{GD}} \;:=\; F_m^{\mathrm{GD}}(\theta^{\mathrm{init}})\,, \quad \text{where} \quad F_m^{\mathrm{GD}}(\psi) \;:=\; \psi - \frac{\eta}{n}\sum\nolimits_{i\le n}\Big(\phi(X_i) - \phi\big(K_{i;\psi}^m(X_i)\big)\Big)\,.$$

The full-batch gradient update with an infinite-length Markov chain is given by

$$\theta_\infty^{\mathrm{GD}} \;:=\; F_\infty^{\mathrm{GD}}(\theta^{\mathrm{init}})\,, \qquad \text{where} \qquad F_\infty^{\mathrm{GD}}(\psi) \;:=\; \psi - \frac{\eta}{n}\sum\nolimits_{i\le n}\Big(\phi(X_i) - \phi(X_1^\psi)\Big)\,.$$

The population gradient update with an infinite-length Markov chain is given by

$$\theta^{\mathrm{pop}} \;:=\; f^{\mathrm{pop}}(\theta^{\mathrm{init}})\,, \qquad \text{where} \quad f^{\mathrm{pop}}(\psi) \;:=\; \psi - \eta\,\mathbb{E}\big[\phi(X_1) - \phi(X_1^\psi)\big]\,,$$

where we use the lowercase f to emphasize that f^{pop} is a deterministic function.

The forthcoming results are summarized below:

$$\theta_{m,B}^{\mathrm{SGDw}} \overset{\text{Lemma E.1}}{\approx} \theta_m^{\mathrm{GD}} \overset{\text{Lemma E.2}}{\approx} \theta_\infty^{\mathrm{GD}} \overset{\text{Lemma E.3}}{\approx} \theta^{\mathrm{pop}} \overset{\text{Lemma E.4}}{\approx} \psi^*\,.$$

Lemma E.1. *Let $\mathcal{F}_n$ be the sigma algebra generated by $\{X_i, K_{i;\theta^{\mathrm{init}}}^m(X_i) \,|\, 1 \le i \le n\}$. Then*

$$\mathbb{E}\big[\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}} \,\big|\, \theta^{\mathrm{init}}, \mathcal{F}_n\big] \;= 0 \qquad \textit{almost surely}\,.$$

Moreover, under A1 and A7, we have

$$\mathbb{E}\big\|\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}\big\|^2 \;\le\; \frac{4\eta^2(\sigma^2 + \kappa_{\nu;m}^2)}{B}\,\mathbb{I}_{\{B<n\}}\,.$$

Proof of Lemma E.1. Write $A := (A_1, \ldots, A_n)$, where

$$A_i \;:=\; \big(\phi(X_i) - \phi\big(K_{i;\theta^{\mathrm{init}}}^m(X_i)\big)\big) - \mathbb{E}\big[\phi(X_1) - \phi\big(K_{i;\theta^{\mathrm{init}}}^m(X_1)\big)\big]\,.$$

Since S^w is uniformly drawn from all size-B subsets of $[n]$ and independently of all other variables, we have that almost surely

$$\mathbb{E}\big[\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}} \,\big|\, \theta^{\mathrm{init}}, \mathcal{F}_n\big] \;=\; \mathbb{E}\Big[\frac{\eta}{B}\sum\nolimits_{i\in S^w} A_i - \frac{\eta}{n}\sum\nolimits_{i\le n} A_i \,\Big|\, A\Big] \;=\; 0\,.$$

To prove the remaining bound, we note that the above relation implies $\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}$ is zero-mean. By the law of total variance, we have

$$\begin{aligned}
\mathbb{E}\big\|\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}\big\|^2 \;&=\; \mathrm{Tr}\,\mathrm{Cov}\big[\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}\big] \;=\; \mathrm{Tr}\,\mathbb{E}\,\mathrm{Cov}\big[\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}} \,\big|\, \theta^{\mathrm{init}}, \mathcal{F}_n\big] \\
&=\; \eta^2\,\mathrm{Tr}\,\mathbb{E}\Big[\mathbb{E}\Big[\Big(\frac{1}{B}\sum\nolimits_{i\in S^w} A_i\Big)\Big(\frac{1}{B}\sum\nolimits_{i\in S^w} A_i\Big)^\top \,\Big|\, A\Big] - \Big(\frac{1}{n}\sum\nolimits_{i\le n} A_i\Big)\Big(\frac{1}{n}\sum\nolimits_{i\le n} A_i\Big)^\top\Big]\,.
\end{aligned}$$

To compute the covariance, recall that S^w is a uniformly drawn size-B subset of $[n]$ with $B = n/N$. Let $\mathcal{P}_N([n])$ be the collection of all partitions of $[n]$ into N size-B subsets. We can generate S^w by the following two-step process:

(i) Uniformly draw a partition $P' = (P_1', \ldots, P_N')$ from $\mathcal{P}_N([n])$;
(ii) Uniformly sample an index K from $[N]$ and set $S^w = P_K'$.

Then we have, almost surely

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{B}\sum\nolimits_{i\in S^w} A_i\Big)\Big(\frac{1}{B}\sum\nolimits_{i\in S^w} A_i\Big)^\top \,\Big|\, A\Big] - \Big(\frac{1}{n}\sum\nolimits_{i\le n} A_i\Big)\Big(\frac{1}{n}\sum\nolimits_{i\le n} A_i\Big)^\top \\
&=\; \frac{1}{|\mathcal{P}_N([n])|}\sum\nolimits_{P'\in\mathcal{P}_N([n])}\Big(\frac{1}{N}\sum\nolimits_{k\le N}\frac{1}{B^2}\sum\nolimits_{i,j\in P_k'} A_iA_j^\top - \frac{1}{n^2}\sum\nolimits_{i,j\le n} A_iA_j^\top\Big) \\
&=\; \frac{1}{|\mathcal{P}_N([n])|}\sum\nolimits_{P'\in\mathcal{P}_N([n])}\Big(\frac{1}{NB^2}\sum\nolimits_{k\le N}\sum\nolimits_{i,j\in P_k'} A_iA_j^\top - \frac{1}{N^2B^2}\sum\nolimits_{k,l\le N}\sum\nolimits_{i\in P_k', j\in P_l'} A_iA_j^\top\Big) \\
&=\; \frac{1}{|\mathcal{P}_N([n])|}\sum\nolimits_{P'\in\mathcal{P}_N([n])}\Big(\frac{N-1}{N^2B^2}\sum\nolimits_{k\le N}\sum\nolimits_{i,j\in P_k'} A_iA_j^\top - \frac{1}{N^2B^2}\sum\nolimits_{k\ne l}\sum\nolimits_{i\in P_k', j\in P_l'} A_iA_j^\top\Big)\,.
\end{aligned}$$

By noting that A_i's are exchangeable, we obtain

$$\begin{aligned}
\mathbb{E}\big\|\theta_{m,B}^{\mathrm{SGDw}}-\theta_m^{\mathrm{GD}}\big\|^2 &= \eta^2 \mathrm{Tr}\,\mathbb{E}\Big[\frac{N-1}{NB}A_1A_1^\top + \frac{(N-1)(B-1)}{NB}A_1A_2^\top - \frac{N-1}{N}A_1A_2^\top\Big] \\
&= \eta^2 \mathrm{Tr}\,\mathbb{E}\Big[\frac{N-1}{NB}A_1A_1^\top - \frac{N-1}{NB}A_1A_2^\top\Big] \\
&= \frac{\eta^2(N-1)}{NB}\big(\mathbb{E}\|A_1\|^2 - \mathbb{E}\langle A_1, A_2\rangle\big) \\
&\overset{(a)}{\le} \frac{2\eta^2}{B}\mathbb{E}\|A_1\|^2 \\
&= \frac{2\eta^2}{B}\mathbb{E}\Big\|\big(\phi(X_1)-\mathbb{E}[\phi(X_1)]\big) - \Big(\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_1)\big) - \mathbb{E}\Big[\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_1)\big)\Big]\Big)\Big\|^2 \\
&\le \frac{4\eta^2}{B}\Big(\mathrm{Tr}\,\mathrm{Cov}[\phi(X_1)] + \mathrm{Tr}\,\mathrm{Cov}\Big[\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_1)\big)\Big]\Big) \\
&\overset{(b)}{\le} \frac{4\eta^2(\sigma^2+\kappa^2_{\nu;m})}{B}\,.
\end{aligned}$$

In (a), we have used a Cauchy-Schwarz inequality; in (b), we have used A1 and A7. Finally we note that if $B = n$, $\theta_{m,B}^{\mathrm{SGDw}} = \theta_m^{\mathrm{GD}}$ almost surely, which implies the desired bound. □

Lemma E.2. *Denote $\mathcal{A}_n$ as the sigma algebra generated by $\theta^{\mathrm{init}}, X_1, \ldots, X_n$. Under A1, A2, A3 and A7, we have that for any $\epsilon > 0$ and $\nu > 2$,*

$$\begin{aligned}
\mathbb{E}\,\big\|\mathbb{E}\big[\theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}}\,\big|\,\mathcal{A}_n\big]\big\|^2 &\le \eta^2\Big(\alpha^m\sigma C_\chi\sqrt{\mathbb{E}\|\theta^{\mathrm{init}}-\psi^*\|^2} + \varepsilon^{\mathrm{init}}_{n,m;\nu}(\epsilon)\Big)^2\,, \\
\mathbb{E}\,\|\theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}}\|^2 &\le \eta^2\Big(\Big(\alpha^m\sigma C_\chi\sqrt{\mathbb{E}\|\theta^{\mathrm{init}}-\psi^*\|^2} + \varepsilon^{\mathrm{init}}_{n,m;\nu}(\epsilon)\Big)^2 + \frac{\kappa^2_{\nu;m}+\sigma^2}{n}\Big)\,.
\end{aligned}$$

Proof of Lemma E.2. The main challenge arises from the possible correlation between θ^{init} and $X_1, \ldots, X_n$. First note that for any $\epsilon > 0$, $\nu > 2$ and a real-valued random variable Y, by Hölder's inequality, we have

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}\big[Y^2\,\mathbb{I}_{\{Y\le\epsilon\}} + Y^2\,\mathbb{I}_{\{Y>\epsilon\}}\big] \\
&\le \epsilon^2 + \mathbb{E}[Y^2\mathbb{I}_{\{Y>\epsilon\}}] \le \epsilon^2 + (\mathbb{E}[Y^\nu])^{2/\nu}\mathbb{P}(Y>\epsilon)^{(\nu-2)/\nu}\,. \qquad (15)
\end{aligned}$$

Also note the useful inequality that for two real-valued random vectors (possibly correlated) V_1, V_2, we have

$$\begin{aligned}
\mathbb{E}[\|V_1+V_2\|^2] \le \mathbb{E}[(\|V_1\|+\|V_2\|)^2] &\le \mathbb{E}\|V_1\|^2 + 2\sqrt{(\mathbb{E}\|V_1\|^2)(\mathbb{E}\|V_2\|^2)} + \mathbb{E}\|V_2\|^2 \\
&= \big(\sqrt{\mathbb{E}\|V_1\|^2} + \sqrt{\mathbb{E}\|V_2\|^2}\big)^2\,. \qquad (16)
\end{aligned}$$

Now to control the first quantity of interest, by using a triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\,\big\|\mathbb{E}\big[\theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}}\,\big|\,\mathcal{A}_n\big]\big\|^2 &= \mathbb{E}\Big\|\,\mathbb{E}\Big[\tfrac{\eta}{n}\textstyle\sum_{i\le n}\big(\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_i)\big) - \phi\big(X_i^{\theta^{\mathrm{init}}}\big)\big)\,\Big|\,\mathcal{A}_n\Big]\Big\|^2 \\
&\overset{(16)}{\le} \eta^2\Big(\sqrt{\mathbb{E}[\Delta_1^2]} + \sqrt{\mathbb{E}[\Delta_2^2]}\Big)^2\,,
\end{aligned}$$

where

$$\begin{aligned}
\Delta_1 &:= \Big\|\,\mathbb{E}\Big[\tfrac{1}{n}\textstyle\sum_{i\le n}\big(\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_i)\big) - \mathbb{E}\big[\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_1')\big)\,\big|\,\theta^{\mathrm{init}}\big]\big)\,\Big|\,\mathcal{A}_n\Big]\Big\| \\
&= \Big\|\,\tfrac{1}{n}\textstyle\sum_{i\le n}\big(\mathbb{E}\big[\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_i)\big)\,\big|\,\theta^{\mathrm{init}}, X_i\big] - \mathbb{E}\big[\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_1')\big)\,\big|\,\theta^{\mathrm{init}}\big]\big)\Big\|\,, \\
\Delta_2 &:= \Big\|\,\mathbb{E}\Big[\tfrac{1}{n}\textstyle\sum_{i\le n}\big(\mathbb{E}\big[\phi\big(K^m_{i;\theta^{\mathrm{init}}}(X_1')\big)\,\big|\,\theta^{\mathrm{init}}\big] - \phi\big(X_i^{\theta^{\mathrm{init}}}\big)\big)\,\Big|\,\mathcal{A}_n\Big]\Big\| \\
&= \Big\|\,\mathbb{E}\big[\phi\big(K^m_{1;\theta^{\mathrm{init}}}(X_1')\big) - \phi\big(X_1^{\theta^{\mathrm{init}}}\big)\,\big|\,\theta^{\mathrm{init}}\big]\Big\|\,,
\end{aligned}$$

and X_1' is an i.i.d. copy of X_1 and in particular independent of $\theta^{\rm init}$. Δ_1 is controlled via (15):

$$
\begin{aligned}
\mathbb{E}[\Delta_1^2] &\le \epsilon^2 + (\mathbb{E}[\Delta_1^\nu])^{2/\nu}\,\mathbb{P}(\Delta_1 > \epsilon)^{\nu/(\nu-2)} \\
&\overset{(a)}{\le} \epsilon^2 + \Big(\mathbb{E}\Big\|\phi(K^m_{1;\theta^{\rm init}}(X_1)) - \mathbb{E}\Big[\phi(K^m_{1;\theta^{\rm init}}(X_1')) \,\Big|\, \theta^{\rm init}\Big]\Big\|^\nu\Big)^{2/\nu} \\
&\qquad \times \mathbb{P}\Bigg(\frac{\big\|\sum_{i\le n}\big(\mathbb{E}\big[\phi\big(K^m_{i;\theta^{\rm init}}(X_i)\big)\,|\,\theta^{\rm init}, X_i\big] - \mathbb{E}\big[\phi\big(K^m_{i;\theta^{\rm init}}(X_1')\big)\,|\,\theta^{\rm init}\big]\big)\big\|}{n} > \epsilon\Bigg)^{\frac{\nu-2}{\nu}} \\
&\overset{(b)}{\le} \epsilon^2 + \kappa^2_{\nu;m}\big(\vartheta^{\rm init}_{n,m}(\epsilon)\big)^{\frac{\nu-2}{\nu}} \\
&= \big(\varepsilon^{\rm init}_{n,m;\nu}(\epsilon)\big)^2 \,. \qquad (17)
\end{aligned}
$$

In (a), we have plugged in the definition of Δ_1 and applied a Jensen's inequality with respect to the empirical average; in (b), we have used A7 to bound the ν-th moment term as

$$
\begin{aligned}
&\Big(\mathbb{E}\Big\|\phi(K^m_{1;\theta^{\rm init}}(X_1)) - \mathbb{E}\Big[\phi(K^m_{1;\theta^{\rm init}}(X_1')) \,\Big|\, \theta^{\rm init}\Big]\Big\|^\nu\Big)^{1/\nu} \\
&= \Big(\mathbb{E}\Big[\mathbb{E}\Big[\Big\|\phi(K^m_{1;\theta^{\rm init}}(X_1)) - \mathbb{E}\Big[\phi(K^m_{1;\theta^{\rm init}}(X_1')) \,\Big|\, \theta^{\rm init}\Big]\Big\|^\nu \,\Big|\, \theta^{\rm init}\Big]\Big]\Big)^{1/\nu} \\
&\le \sup_{\psi\in\Psi}\Big(\mathbb{E}\Big[\Big\|\phi(K^m_{1;\psi}(X_1)) - \mathbb{E}\big[\phi(K^m_{1;\psi}(X_1'))\big]\Big\|^\nu\Big]\Big)^{1/\nu} \\
&= \sup_{\psi\in\Psi}\Big(\mathbb{E}\Big[\Big\|\phi(K^m_{1;\psi}(X_1)) - \mathbb{E}\big[\phi(K^m_{1;\psi}(X_1))\big]\Big\|^\nu\Big]\Big)^{1/\nu} \;\le\; \kappa_{\nu;m}
\end{aligned}
$$

and recalled the definitions of $\vartheta^{\rm init}_{n,m}$ and $\varepsilon^{\rm init}_{n,m;\nu}$. On the other hand,

$$
\begin{aligned}
\mathbb{E}[\Delta_2^2] \;&=\; \mathbb{E}\Big\|\int_{\mathbb{R}^d}\phi(x)(K^m_{1;\theta^{\rm init}}p_{\psi^*})(x)dx - \mathbb{E}\big[\phi(X_1^{\theta^{\rm init}})\big|\theta^{\rm init}\big]\Big\|^2 \\
&\overset{(a)}{=} \mathbb{E}\Big\|\int_{\mathbb{R}^d}(K^m_{1;\theta^{\rm init}}\phi)(x)\,p_{\psi^*}(x)dx - \mathbb{E}_{p_{\theta^{\rm init}}}[\,\phi\,]\Big\|^2 \\
&\overset{(b)}{=} \mathbb{E}\Big\|\int_{\mathbb{R}^d}\big(K^m_{1;\theta^{\rm init}}\big(\phi - \mathbb{E}_{p_{\theta^{\rm init}}}[\,\phi\,]\big)\big)(x)\times p_{\psi^*}(x)dx\Big\|^2 \\
&\overset{(c)}{=} \mathbb{E}\Big\|\int_{\mathbb{R}^d}\big(K^m_{1;\theta^{\rm init}}\big(\phi - \mathbb{E}_{p_{\theta^{\rm init}}}[\,\phi\,]\big)\big)(x)\times (p_{\psi^*}(x) - p_{\theta^{\rm init}}(x))dx\Big\|^2 \\
&\overset{(d)}{=} \textstyle\sum_{l=1}^d \mathbb{E}\Big(\int_{\mathbb{R}^d}\big(K^m_{t1;\theta^{\rm init}}\big(\phi - \mathbb{E}_{p_{\theta^{\rm init}}}[\,\phi\,]\big)\big)(x)^\top e_l \times p_{\theta^{\rm init}}(x)\times\frac{p_{\psi^*}(x) - p_{\theta^{\rm init}}(x)}{p_{\theta^{\rm init}(x)}}\,dx\Big)^2 \\
&\overset{(e)}{\le} \textstyle\sum_{l=1}^d \mathbb{E}\Big[\Big(\int_{\mathbb{R}^d}\big(\big(K^m_{t1;\theta^{\rm init}}\big(\phi - \mathbb{E}_{p_{\theta^{\rm init}}}[\,\phi\,]\big)\big)(x)^\top e_l\big)^2 p_{\theta^{\rm init}}(x)dx \\
&\qquad\qquad \times\int_{\mathbb{R}^d}\Big(\frac{p_{\psi^*}(x) - p_{\theta^{\rm init}}(x)}{p_{\theta^{\rm init}(x)}}\Big)^2 p_{\theta^{\rm init}}(x)dx\Big)^2\Big] \\
&\overset{(f)}{\le} \textstyle\sum_{l=1}^d \mathbb{E}\Big(\alpha^{2m}\,\mathrm{Tr}\,\mathrm{Cov}\big[\phi(X_1^{\theta^{\rm init}})\,\big|\,\theta^{\rm init}\big]\,\chi^2(p_{\psi^*}, p_{\theta^{\rm init}})\Big)^2 \\
&\overset{(g)}{\le} \alpha^{2m}\sigma^2 C_\chi^2\,\mathbb{E}\|\theta^{\rm init} - \psi^*\|^2 \,.
\end{aligned}
$$

In (a), we have used that $(Kf)(x) = \int K(x,y)f(y)dy$; in (b), we have used that the Markov operator leaves the constant function invariant; in (c), we used that $K_{1;\theta^{\rm init}}$ leaves $p_{\theta^{\rm init}}$ invariant; in (d), we denoted $(e_l)_{l\le d}$ as the standard basis vectors of $\mathbb{R}^d$ and multiplied and divided by $p_{\theta^{\rm init}}(x)$; in (e), we have used a Cauchy-Schwarz inequality; in (f), we have used the definition of the spectral gap α in A3; in (g), we have used A1 and A2. Combining the bounds gives the first inequality that

$$
\mathbb{E}\,\big\|\mathbb{E}\big[\theta^{\rm GD}_m - \theta^{\rm GD}_\infty\,\big|\,\mathcal{A}_n\big]\big\|^2 \;\le\; \eta^2\Big(\alpha^m\sigma C_\chi\,\sqrt{\mathbb{E}\|\theta^{\rm init} - \psi^*\|^2} \;+\; \varepsilon^{\rm init}_{n,m;\nu}(\epsilon)\Big)^2 \,.
$$

Now we handle the second quantity by conditioning on $\theta^{\rm init}$ and perform a bias-variance decomposition:

$$
\begin{aligned}
\mathbb{E}\,\|\theta^{\rm GD}_m - \theta^{\rm GD}_\infty\|^2 \;&= \eta^2\,\mathbb{E}\Big[\,\mathbb{E}\Big[\Big\|\frac{1}{n}\textstyle\sum_{i\le n}\big(\phi\big(K^m_{i;\theta^{\rm init}}(X_i)\big) - \phi(X_i^{\theta^{\rm init}})\big)\Big\|^2\,\Big|\,\mathcal{A}_n\Big]\,\Big] \\
&= \eta^2(Q_B + Q_V)\,,
\end{aligned}
$$

where

$$Q_B := \mathbb{E}\Big\|\mathbb{E}\Big[\frac{1}{n}\sum_{i\le n}\big(\phi\big(K^m_{i;\theta^{\rm init}}(X_i)\big)-\phi(X_i^{\theta^{\rm init}})\big)\,\Big|\,\mathcal{A}_n\Big]\Big\|^2 ,$$

$$Q_V := \mathbb{E}\Big[\mathrm{Tr}\Big(\mathrm{Cov}\Big[\frac{1}{n}\sum_{i\le n}\phi\big(K^m_{i;\theta^{\rm init}}(X_i)\big)\Big|\mathcal{A}_n\Big]+\mathrm{Cov}\Big[\frac{1}{n}\sum_{i\le n}\phi(X_i^{\theta^{\rm init}})\Big|\theta^{\rm init}\Big]\Big)\Big] .$$

Note that the covariance terms separate because $X_i^{\theta^{\rm init}}$ is independent of $K^m_{i;\theta^{\rm init}}(X_i)$ conditioning on $\theta^{\rm init}$. $\eta^2 Q_B$ is exactly the quantity controlled above, so it suffices to bound the variance term Q_V. By explicitly computing the second covariance term while noting that $X_1^{\theta^{\rm init}},\dots,X_n^{\theta^{\rm init}}$ are conditionally i.i.d. given $\theta^{\rm init}$ and $K^m_{i;\theta^{\rm init}}(X_i)$'s are conditionally independent across $1\le i\le n$ given $\mathcal{A}_n$, we have

$$\begin{aligned}
Q_V &= \frac{\sum_{i\le n}\mathbb{E}[\mathrm{Tr}\,\mathrm{Cov}[\phi(K^m_{i;\theta^{\rm init}}(X_i))\,|\,\theta^{\rm init},X_i]]}{n^2}+\frac{\mathbb{E}\big[\mathrm{Tr}\,\mathrm{Cov}[\phi(X_1^{\theta^{\rm init}})|\theta^{\rm init}]\big]}{n}\\
&\overset{(a)}{=} \frac{\mathbb{E}[\mathrm{Tr}\,\mathrm{Cov}[\phi(K^m_{1;\theta^{\rm init}}(X_1))\,|\,\theta^{\rm init},X_1]]}{n}+\frac{\mathbb{E}\big[\mathrm{Tr}\,\mathrm{Cov}[\phi(X_1^{\theta^{\rm init}})|\theta^{\rm init}]\big]}{n}\\
&\le \frac{\kappa^2_{\nu;m}+\sigma^2}{n} ,
\end{aligned}$$

where we have used A4, A7 and A1 in the last line. Combining the bounds, we obtain that

$$\begin{aligned}
\mathbb{E}\,\|\theta^{\rm GD}_m-\theta^{\rm GD}_\infty\|^2 &= \eta^2(Q_B+Q_V)\\
&\le \eta^2\Big(\Big(\alpha^m\sigma C_\chi\sqrt{\mathbb{E}\|\theta^{\rm init}-\psi^*\|^2}+\varepsilon^{\rm init}_{n,m;\nu}(\epsilon)\Big)^2+\frac{\kappa^2_{\nu;m}+\sigma^2}{n}\Big) .
\end{aligned}$$

□

Lemma E.3. *Under A1,* $\mathbb{E}\,\|\theta^{\rm GD}_\infty-\theta^{\rm pop}\|^2 \le \frac{4\eta^2\sigma^2}{n}$.

Proof of Lemma E.3. Since both $F^{\rm GD}_\infty$ and $f^{\rm pop}$ involve infinite-length Markov chains, the initializations do not matter and we can decouple the stochasticity of X_i and $K_{i;\theta^{\rm init}}$. In particular,

$$\begin{aligned}
\mathbb{E}\,\|\theta^{\rm GD}_\infty-\theta^{\rm pop}\|^2 &= \eta^2\,\mathbb{E}\Big\|\frac{1}{n}\sum_{i\le n}\big(\phi(X_i)-\phi\big(X_i^{\theta^{\rm init}}\big)\big)-\mathbb{E}\big[\phi(X_1)-\phi\big(X_1^{\theta^{\rm init}}\big)\big]\Big\|^2\\
&\le \eta^2\big(Q_1'+2\sqrt{Q_1'Q_2'}+Q_2'\big) ,
\end{aligned}$$

where

$$\begin{aligned}
Q_1' &:= \mathbb{E}\Big\|\frac{1}{n}\sum_{i\le n}\big(\phi(X_i)-\mathbb{E}[\phi(X_1)]\big)\Big\|^2 = \frac{\mathrm{Tr}\,\mathrm{Cov}[\phi(X_1)]}{n} = \frac{\mathrm{Tr}\,\nabla^2_\theta\log Z(\psi^*)}{n} \le \frac{\sigma^2}{n} ,\\
Q_2' &:= \mathbb{E}\Big\|\frac{1}{n}\sum_{i\le n}\big(\phi(X_i^{\theta^{\rm init}})-\mathbb{E}\big[\phi\big(X_1^{\theta^{\rm init}}\big)\big]\big)\Big\|^2\\
&= \frac{\mathbb{E}\big[\mathrm{Tr}\,\mathrm{Cov}[\phi(X_1^{\theta^{\rm init}})|\theta^{\rm init}]\big]}{n} = \frac{\mathbb{E}\big[\mathrm{Tr}\,\nabla^2_\theta\log Z(\theta^{\rm init})\big]}{n} \le \frac{\sigma^2}{n} .
\end{aligned}$$

In the computations above, we have used the relation $\nabla^2_\theta\log Z(\theta) = \mathrm{Cov}_{X\sim p_\theta}[\phi(X)]$ and the assumption $\sup_{\theta\in\Psi}\mathrm{tr}(\nabla^2_\theta\log Z(\theta))=\sigma^2$ from A1. This implies the desired bound. □

Lemma E.4. *Under A1,* $\mathbb{E}\,\|\theta^{\rm pop}-\psi^*\|^2 \le \big(1-2\mu\eta+L^2\eta^2\big)\,\mathbb{E}\,\|\theta^{\rm init}-\psi^*\|^2$.

Proof of Lemma E.4. Recall that

$$f^{\rm pop}(\theta') = \theta-\eta\,\mathbb{E}\big[\phi(X_1)-\phi(X_1^{\theta'})\big] = \theta-\eta\,\big(\nabla_\psi\log Z(\psi^*)-\nabla_\psi\log Z(\theta')\big) .$$

By construction, $f^{\rm pop}$ is deterministic and $f^{\rm pop}(\psi^*)=\psi^*$. By plugging in the recursions and expanding the square, we get that

$$\begin{aligned}
\mathbb{E}\,\big\|\theta^{\rm pop}-\psi^*\big\|^2 &= \mathbb{E}\,\big\|f^{\rm pop}(\theta^{\rm init})-f^{\rm pop}(\psi^*)\big\|^2\\
&= \mathbb{E}\,\big\|(\theta^{\rm init}-\psi^*)-\eta(\nabla_\psi\log Z(\theta^{\rm init})-\nabla_\psi\log Z(\psi^*))\big\|^2\\
&= \mathbb{E}\,\big\|\theta^{\rm init}-\psi^*\big\|^2-2\eta\,\mathbb{E}\big[\big\langle\theta^{\rm init}-\psi^*\,,\,\nabla_\psi\log Z(\theta^{\rm init})-\nabla_\psi\log Z(\psi^*)\big\rangle\big]\\
&\quad+\eta^2\,\mathbb{E}\,\big\|\nabla_\psi\log Z(\theta^{\rm init})-\nabla_\psi\log Z(\psi^*)\big\|^2\\
&\le \mathbb{E}\,\big\|\theta^{\rm init}-\psi^*\big\|^2-2\mu\eta\,\mathbb{E}\,\big\|\theta^{\rm init}-\psi^*\big\|^2+L^2\eta^2\,\mathbb{E}\,\big\|\theta^{\rm init}-\psi^*\big\|^2 .
\end{aligned}$$

In the last line, we have recalled $\inf_{\psi\in\Psi}\lambda_{\min}(\nabla^2_\psi \log Z(\psi)) = \mu$ and $\sup_{\theta\in\Psi}\lambda_{\max}(\nabla^2_\psi \log Z(\psi)) = L$ by A1 and applied Lemma C.8. Combining the coefficients gives the desired statement. □

F Proofs for offline SGD

We prove Theorem B.1 (which directly implies Theorem 4.3) and Theorem B.2 in this section. The key ingredient of both proofs is Lemma F.1 below, which provides an iterative error bound for the SGD-with-replacement scheme by combining different approximation bounds in Appendix E. Throughout this section, we denote $\delta^{\mathrm{SGDw}}_{t,j} := \mathbb{E}\left\|\psi^{\mathrm{SGDw}}_{t,j} - \psi^*\right\|^2$.

Lemma F.1. *Under A1, A2, A3, A4 and A7, we have that for* $1 \le j \le N-1$,

$$\sqrt{\delta^{\mathrm{SGDw}}_{t,j}} \;\le\; \Big(1 - \eta_t\Big(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2}\eta_t\Big)\Big)\sqrt{\delta^{\mathrm{SGDw}}_{t,j-1}} \;+\; \eta_t\Big(\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}}\Big)\,,$$

where $1 - \eta_t\big(\mu - \alpha^m\sigma C_\chi - \frac{L^2}{2}\eta_t\big) > 0\,.$

Proof of Lemma F.1. We first remark that in view of Lemma C.4, the projection step in Algorithm 2 does not increase $\delta^{\mathrm{SGDw}}_{t,j}$, so it suffices to bound $\delta^{\mathrm{SGDw}}_{t,j}$ as if projection is not performed on $\psi^{\mathrm{SGDw}}_{t,j}$. To apply the results from Appendix E, we identify $\theta^{\mathrm{init}} = \psi^{\mathrm{SGDw}}_{t,j-1}$ and $\eta = \eta_t$, which allows us to write $\psi^{\mathrm{SGDw}}_{t,j} = \theta^{\mathrm{SGDw}}_{m,B}$. This also implies $\mathbb{E}\|\theta^{\mathrm{init}} - \psi^*\|^2 = \delta^{\mathrm{SGDw}}_{t,j-1}$ and $\varepsilon^{\mathrm{init}}_{n,m;\nu}(\epsilon) \le \varepsilon^{\mathrm{SGDw}}_{\nu;n,m,t}(\epsilon)$.

By adding and subtracting the auxiliary gradient updates followed by expanding the square, we obtain

$$
\begin{aligned}
\delta_{t,j}^{\mathrm{SGDw}} &= \mathbb{E}\left\|\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}} + \theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}} + \theta_\infty^{\mathrm{GD}} - \theta^{\mathrm{pop}} + \theta^{\mathrm{pop}} - \psi^*\right\|^2 \\
&\overset{(a)}{=} \mathbb{E}\left\|\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}\right\|^2 + \mathbb{E}\left\|\theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}}\right\|^2 + \mathbb{E}\left\|\theta_\infty^{\mathrm{GD}} - \theta^{\mathrm{pop}}\right\|^2 + \mathbb{E}\left\|\theta^{\mathrm{pop}} - \psi^*\right\|^2 \\
&\quad + 2\,\mathbb{E}\left\langle \theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}\,,\, \theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}}\right\rangle + 2\,\mathbb{E}\left\langle \theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}}\,,\, \theta_\infty^{\mathrm{GD}} - \theta^{\mathrm{pop}}\right\rangle \\
&\quad + 2\,\mathbb{E}\left\langle \mathbb{E}\left[\theta_{m,B}^{\mathrm{SGDw}} - \theta_m^{\mathrm{GD}} \,\middle|\, \theta^{\mathrm{init}}, \mathcal{F}_n\right]\,,\, \theta^{\mathrm{pop}} - \psi^*\right\rangle + 2\,\mathbb{E}\left\langle \theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}}\,,\, \theta_\infty^{\mathrm{GD}} - \theta^{\mathrm{pop}}\right\rangle \\
&\quad + 2\,\mathbb{E}\left\langle \mathbb{E}\left[\theta_m^{\mathrm{GD}} - \theta_\infty^{\mathrm{GD}} \,\middle|\, \mathcal{A}_n\right]\,,\, \theta^{\mathrm{pop}} - \psi^*\right\rangle + 2\,\mathbb{E}\left\langle \theta_\infty^{\mathrm{GD}} - \theta^{\mathrm{pop}}\,,\, \theta^{\mathrm{pop}} - \psi^*\right\rangle \\
&\overset{(b)}{\le} \frac{4\eta_t^2(\sigma^2+\kappa_{\nu;m}^2)}{\sqrt{B}}\mathbb{I}_{\{B<n\}} + \eta_t^2\Big(\Big(\alpha^m\sigma C_\chi\sqrt{\delta_{t,j-1}^{\mathrm{SGDw}}} + \varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon)\Big)^2 + \frac{\kappa_{\nu;m}^2+\sigma^2}{n}\Big) \\
&\quad + \frac{4\eta_t^2\sigma^2}{n} + (1-2\mu\eta_t + L^2\eta_t^2)\delta_{t,j-1}^{\mathrm{SGDw}} \\
&\quad + 2\frac{2\eta_t\sqrt{\sigma^2+\kappa_{\nu;m}^2}}{\sqrt{B}}\mathbb{I}_{\{B<n\}}\eta_t\Big(\alpha^m\sigma C_\chi\sqrt{\delta_{t,j-1}^{\mathrm{SGDw}}} + \varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon) + \frac{\sqrt{\kappa_{\nu;m}^2+\sigma^2}}{\sqrt{n}}\Big) \\
&\quad + 2\frac{2\eta_t\sqrt{\sigma^2+\kappa_{\nu;m}^2}}{\sqrt{B}}\mathbb{I}_{\{B<n\}}\frac{2\eta_t\sigma}{\sqrt{n}} + 0 \\
&\quad + 2\eta_t\Big(\alpha^m\sigma C_\chi\sqrt{\delta_{t,j-1}^{\mathrm{SGDw}}} + \varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon) + \frac{\sqrt{\kappa_{\nu;m}^2+\sigma^2}}{\sqrt{n}}\Big)\frac{2\eta_t\sigma}{\sqrt{n}} \\
&\quad + 2\eta_t\Big(\alpha^m\sigma C_\chi\sqrt{\delta_{t,j-1}^{\mathrm{SGDw}}} + \varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon)\Big)\sqrt{(1-2\mu\eta_t+L^2\eta_t^2)\delta_{t,j-1}^{\mathrm{SGDw}}} \\
&\quad + 2\frac{2\eta_t\sigma}{\sqrt{n}}\sqrt{(1-2\mu\eta_t+L^2\eta_t^2)\delta_{t,j-1}^{\mathrm{SGDw}}} \\
&\overset{(c)}{\le} \Big(1-2\mu\eta_t + L\eta_t^2 + \eta_t^2\alpha^{2m}\sigma^2C_\chi^2 + 2\eta_t\alpha^m\sigma C_\chi\sqrt{1-2\mu\eta_t+L^2\eta_t^2}\Big)\times\delta_{t,j-1}^{\mathrm{SGDw}} \\
&\quad + 2\Bigg(\eta_t^2\alpha^m\sigma C_\chi\varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon) + \frac{2\eta_t^2(\sqrt{\sigma^2+\kappa_{\nu;m}^2}+\sigma)\,\alpha^m\sigma C_\chi}{\sqrt{B}} \\
&\qquad + \eta_t\varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon)\sqrt{1-2\mu\eta_t+L_2\eta_t^2} + \frac{2\eta_t\sigma\sqrt{1-2\mu\eta_t+L^2\eta_t^2}}{\sqrt{n}}\Bigg)\times\sqrt{\delta_{t,j-1}^{\mathrm{SGDw}}} \\
&\quad + \eta_t^2\Big(\frac{4(\sigma^2+\kappa_{\nu;m}^2)}{B} + \big(\varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon)\big)^2 + \frac{5\kappa_{\nu;m}^2+9\sigma^2}{B} + \frac{4\sqrt{\sigma^2+\kappa_{\nu;m}^2}\,\varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon)}{\sqrt{B}} \\
&\qquad + \frac{8\sigma\sqrt{\sigma^2+\kappa_{\nu;m}^2}}{B} + \frac{4\sigma\varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon)}{\sqrt{B}}\Big) \\
&=: (A_1)\times\delta_{t,j-1}^{\mathrm{SGDw}} + (A_2)\times\sqrt{\delta_{t,j-1}^{\mathrm{SGDw}}} + (A_3)\,.
\end{aligned}
$$

In (a), we have expanded the square, used $\mathcal{F}_n$ defined in Lemma E.1 and $\mathcal{A}_n$ defined in Lemma E.2, and noted that $\theta^{\mathrm{pop}} - \psi^*$ is almost surely constant given θ^{init}; in (b), we have applied Lemmas E.1, E.2, E.3 and E.4 under A1, A2, A3, A4 and A7, and used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$; in (c), we have grouped the terms by powers of $\delta_{t,j-1}^{\mathrm{SGDw}}$, bounded the indicator function from above by 1 and used $B \le n$ to replace n in the denominator by B. While the computation above is complicated, we remark that the key steps are taking the conditional expectations for the cross-terms in (b), which gives us a tighter bound than directly applying a triangle inequality of the form $\mathbb{E}\|Y_1 + Y_2\|^2 \le (\sqrt{\mathbb{E}\|Y_1\|^2} + \sqrt{\mathbb{E}\|Y_2\|^2})$. To further simplify the bounds, we seek to bound each

coefficient by a square, which yields

$$
\begin{aligned}
(A_1) &= \Big(\eta_t \alpha^m \sigma C_\chi + \sqrt{1 - 2\mu\eta_t + L\eta_t^2}\Big)^2 , \\
(A_2) &\le 2\Big(\eta_t \alpha^m \sigma C_\chi + \sqrt{1 - 2\mu\eta_t + L\eta_t^2}\Big) \times \eta_t \Big(\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{4\sigma + 2\kappa_{\nu;m}}{\sqrt{B}}\Big) , \\
(A_3) &\le \eta_t^2 \Big(\big(\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon)\big)^2 + \frac{8\sigma^2 + 4\kappa^2_{\nu;m}}{\sqrt{B}} \varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{21\sigma^2 + 17\kappa^2_{\nu;m}}{B} \Big) \\
&\le \eta_t^2 \Big(\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}}\Big)^2 .
\end{aligned}
$$

This implies

$$
\begin{aligned}
\delta^{\mathrm{SGDw}}_{t,j} &\le (A_1) \times \delta^{\mathrm{SGDw}}_{t,j-1} + (A_2) \times \sqrt{\delta^{\mathrm{SGDw}}_{t,j-1}} + (A_3) \\
&\le \Big(\Big(\eta_t \alpha^m \sigma C_\chi + \sqrt{1 - 2\mu\eta_t + L\eta_t^2}\Big) \sqrt{\delta^{\mathrm{SGDw}}_{t,j-1}} + \eta_t \Big(\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}}\Big)\Big)^2 .
\end{aligned}
$$

Now note that since $\mu \le L$ by the definitions in A1, $2\mu\eta_t - L^2\eta_t^2 \le 2L\eta_t - L^2\eta_t^2 \le 1$. By using $\sqrt{1-x} \le 1 - \frac{x}{2}$ for all $x \le 1$, we get that

$$
\sqrt{1 - 2\mu\eta_t + L^2\eta_t^2} \;\le\; 1 - \mu\eta_t + \frac{L^2}{2}\eta_t^2 .
$$

Substituting this into the earlier bound and taking a square-root, we obtain the desired bound that

$$
\sqrt{\delta^{\mathrm{SGDw}}_{t,j}} \;\le\; \Big(1 - \eta_t\Big(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2}\eta_t\Big)\Big) \sqrt{\delta^{\mathrm{SGDw}}_{t,j-1}} \;+\; \eta_t \Big(\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}}\Big) .
$$

Moreover, since $L \ge \mu$ by definition from A1, we have

$$
1 - \eta_t\Big(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2}\eta_t\Big) \;=\; \Big(\frac{L}{\sqrt{2}}\eta_t - 1\Big)^2 + (\sqrt{2}\,L - \mu)\eta_t + \alpha^m \sigma C_\chi \eta_t \;>\; 0 ,
$$

which finishes the proof. □

F.1 Proof of Theorem B.1

Under A1, A2, A3, A4 and A7, Lemma F.1 implies

$$
\sqrt{\delta^{\mathrm{SGDw}}_{t,j}} \;\le\; \Big(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta}\Big) \sqrt{\delta^{\mathrm{SGDw}}_{t,j-1}} \;+\; C t^{-\beta}\, \sigma^{\mathrm{SGDw}}_{n,T}
$$

for $1 \le t \le T$ and $1 \le j \le N$, where we have used $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi$, $\eta_t = C t^{-\beta}$ and

$$
\varepsilon^{\mathrm{SGDw}}_{n,m,t;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}} \;\le\; \varepsilon^{\mathrm{SGDw}}_{n,m,T;\nu}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}} \;=\; \sigma^{\mathrm{SGDw}}_{n,T} .
$$

By an induction on $j = 1, \ldots, N$, we have

$$
\begin{aligned}
\sqrt{\delta^{\mathrm{SGDw}}_{t,N}} \;&\le \Big(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta}\Big)^N \sqrt{\delta^{\mathrm{SGDw}}_{t,0}} \\
&\quad + C \sigma^{\mathrm{SGDw}}_{n,T}\, t^{-\beta} \textstyle\sum_{j=1}^N \Big(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta}\Big)^{j-1} .
\end{aligned}
$$

By noting that $\delta^{\mathrm{SGDw}}_{t,0} = \delta^{\mathrm{SGDw}}_{t-1,N}$ almost surely for $t \ge 2$ and using another induction on $t = 1, \ldots, T$,

$$
\sqrt{\delta^{\mathrm{SGDw}}_{T,N}} \le Q_0 \sqrt{\delta^{\mathrm{SGDw}}_{0,0}} \;+\; C \sigma^{\mathrm{SGDw}}_{n,T} \textstyle\sum_{t=1}^T Q_t\, A_t\, t^{-\beta} , \tag{18}
$$

where

$$
Q_t \coloneqq \textstyle\prod_{s=t+1}^T \Big(1 - \tilde{\mu}_m C s^{-\beta} + \frac{L^2 C^2}{2} s^{-2\beta}\Big)^N \;\text{ and }\; A_t \coloneqq \textstyle\sum_{j=1}^N \Big(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta}\Big)^{j-1} .
$$

Note that by Lemma F.1, we also have that $1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta} > 0$. This allows us to apply Lemma C.3 to obtain

$$
\kappa_0 \;\le\; \exp\Big(1 - N \tilde{\mu}_m C \varphi_{1-\beta}(T+1) + \frac{N L^2 C^2}{2} \varphi_{1-2\beta}(T+1)\Big) \;=\; E_1^{T,N} .
$$

To control $\sum_{t=1}^{T} Q_t A_t t^{-\beta}$, recall that $\beta \in [0,1]$, $\tilde{\mu}_m > 0$, $\frac{L^2C^2}{2} > 0$, and that

$$E_2^{T,N} \;=\; \exp\Big(-\frac{N\tilde{\mu}_m C}{2}\varphi_{1-\beta}(T+1) + 2NL^2C^2\varphi_{1-2\beta}(T+1)\Big)\,.$$

We can now apply Lemma C.3: If $\beta \notin \{\frac{1}{2}, 1\}$, then

$$\sum_{t=1}^{T} Q_t\, A_t\, t^{-\beta} \;\le\; \frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)}\frac{N}{(T+1)^{\beta}}} + \frac{3^{\beta}(1+\tilde{\mu}_m C)^{N-1}(T+2)^{\beta}}{L^2C^2}\, E_2^{T,N}\,.$$

If $\beta = \frac{1}{2}$, i.e. $2\beta = 1$, we have

$$\sum_{t=1}^{T} Q_t\, A_t\, t^{-\beta} \;\le\; \frac{4}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}} + 2N(1+\tilde{\mu}_m C)^{N-1}\varphi_{\frac{1}{2}-L^2C^2N}(T+1)\, E_2^{T,N}\,.$$

If $\beta = 1$, we get that

$$\sum_{t=1}^{T} Q_t\, A_t\, t^{-\beta} \;\le\; \frac{4}{\tilde{\mu}_m C} + \frac{3N\big(1+\frac{L^2C^2}{2}\big)^{N-1} e^{2L^2C^2N}\,\log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}}$$

Substituting the bounds into (18), we get the desired bound that $\sqrt{\delta_{T,N}^{\mathrm{SGDw}}}$ is upper bounded by

$$\begin{cases} E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} \;+\; C\sigma_{n,T}^{\mathrm{SGDw}}\Big(\frac{4e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}}}{\tilde{\mu}_m C} + 2N(1+\tilde{\mu}_m C)^{N-1}\varphi_{\frac{1}{2}-L^2C^2N}(T+1)\, E_2^{T,N}\Big) & \text{for } \beta = \frac{1}{2}\,, \\ E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} + C\sigma_{n,T}^{\mathrm{SGDw}}\Big(\frac{4}{\tilde{\mu}_m C} + \frac{3N\big(1+\frac{L^2C^2}{2}\big)^{N-1} e^{2L^2C^2N}\,\log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}}\Big) & \text{for } \beta = 1\,, \\ E_1^{T,N}\sqrt{\delta_{0,0}^{\mathrm{SGDw}}} \;+\; C\sigma_{n,T}^{\mathrm{SGDw}}\Big(\frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)}\frac{N}{(T+1)^{\beta}}} + \frac{3^{\beta}(1+\tilde{\mu}_m C)^{N-1}(T+2)^{\beta}}{L^2C^2}\, E_2^{T,N}\Big) & \text{otherwise}\,. \end{cases}$$

□

F.2 Proof of Theorem B.2

Recall that $\delta_{t,j}^{\mathrm{SGDo}} := \mathbb{E}\left\|\psi_{t,j}^{\mathrm{SGDo}} - \psi^*\right\|^2$. To apply the results from Appendix E to $\psi_{t,j}^{\mathrm{SGDo}}$, we condition on S_t^o, which in particular fixes $S_{t,j}^o$, the last size-B subset of $[n]$ chosen. We then identify $\theta^{\mathrm{init}} = \psi_{t,j-1}^{\mathrm{SGDo}}$, $\eta = \eta_t$ and the dataset used as $\mathcal{D}_{t,j}^o := (X_i \,:\, i \in S_{t,j}^o)$, which allows us to identify $\psi_{t,j}^{\mathrm{SGDo}} = \theta_m^{\mathrm{GD}}$, the full-batch gradient descent update using $\mathcal{D}_{t,j}^o$. Meanwhile, we note that almost surely $\theta_m^{\mathrm{GD}} = \theta_{m,B}^{\mathrm{SGDw}}$, the SGD-with-replacement iterate that uses the full dataset $\mathcal{D}_{t,j}^o$. Observe that the proof of Lemma F.1 holds with $\delta_{t,j-1}^{\mathrm{SGDo}}$ replaced by any random initialization θ^{init} possibly correlated with $X_1, \ldots, X_n$, which allows us to obtain

$$\sqrt{\delta_{t,j}^{\mathrm{SGDo}}} \;\le\; \Big(1 - \eta_t\Big(\mu - \alpha^m\sigma C_\chi - \frac{L^2}{2}\eta_t\Big)\Big)\sqrt{\delta_{t,j-1}^{\mathrm{SGDo}}} \;+\; \eta_t\Big(\varepsilon_{n,m,t;\nu}^{\mathrm{SGDw}}(\epsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}}\Big)\,.$$

Since the error recursion for $\delta_{t,j}^{\mathrm{SGDo}}$ is identical to that of $\delta_{t,j}^{\mathrm{SGDw}}$ in Lemma F.1, the proof of Theorem 4.3 follows directly, thereby yielding an identical result for $\delta_{T,N}^{\mathrm{SGDo}}$ as with $\delta_{T,N}^{\mathrm{SGDw}}$ in Theorem 4.3. □

G Proofs for tail probability bounds in offline SGD

We present the proofs for results in Appendix B.3 that control the tail probability terms $\vartheta_{\nu;n,m,T}^{\mathrm{SGDw}}$ and $\varepsilon_{\nu;n,m,T}^{\mathrm{SGDw}}$.

Proof of Lemma B.3. For $\delta > 0$, let N_δ be the δ-covering number of Ψ, which satisfies $N_\delta \le \big(r_\Psi(1+2/\delta)\big)^p$ (Example 5.8, [50]). Note also that by the Jensen's inequality applied to $\mathbb{E}[\,\bullet\,|X_1]$ and Assumption A6, there exist some $\sigma_m, \zeta_m > 0$ such that, for any $z \in \mathbb{R}^p$ with $\|z\| \le \zeta_m$,

$$\mathbb{E}[e^{z^\top(\mathbb{E}[\phi(K_{\psi^*}^m(X_1))|X_1] - \mathbb{E}[\phi(K_{\psi^*}^m(X_1))])}] \;\le \mathbb{E}[e^{z^\top(\phi(K_{\psi^*}^m(X_1)) - \mathbb{E}[\phi(K_{\psi^*}^m(X_1))])}] \;\le\; e^{\sigma_m^2\|z\|^2/2}\,.$$

Meanwhile under A5, recycling the proof of Lemma 3.1 of [21] shows that if $C_m\delta/\sqrt{p} < \sigma_m^2\zeta_m$,

$$\mathbb{P}\Big(\sup_{\psi\in\Psi}\Big\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\phi\big(K_\psi^m(X_i)\big)\big|X_i\big] - \mathbb{E}\big[\phi(K_\psi^m(X_i))\big]\Big\| > 3C_m\delta\Big) \;\le\; 2N_\delta p\exp\Big(-\frac{nC_m^2\delta^2}{2p\sigma_m^2}\Big)$$
$$\le 2(r_\Psi)^p\exp\Big(p\log(1+2\delta^{-1}) - \frac{nC_m^2\delta^2}{2p\sigma_m^2}\Big)\,.$$

Since the probability above is an upper bound to $\vartheta_{n,m,T}^{\mathrm{SGDw}}(3C_m\delta)$, we get that if $C_m\delta/\sqrt{p} < \sigma_m^2\zeta_m$,

$$\Big(\varepsilon_{\nu;n,m,T}^{\mathrm{SGDw}}(3C_m\delta)\Big)^2 \;=\; 9C_m^2\delta^2 + \kappa_{\nu;m}^2\Big(\vartheta_{n,m,T}^{\mathrm{SGDw}}\Big(\frac{C_m\delta}{\sqrt{p}}\Big)\Big)^{\frac{\nu-2}{\nu}}$$
$$\le 9C_m^2\delta^2 + \kappa_{\nu;m}^2 2^{\frac{2(\nu-2)}{\nu}}(r_\Psi)^{\frac{(\nu-2)p}{\nu}}\exp\Big(\frac{(\nu-2)p}{\nu}\log(1+2\delta^{-1}) - \frac{n(\nu-2)C_m^2\delta^2}{2\nu p\sigma_m^2}\Big)\,.$$

Recall that by assumption, $\frac{\log n}{n} < \frac{\sigma_m^2\zeta_m^2}{p+\nu-2}$. We now choose

$$\delta \;=\; \frac{\sigma_m}{C_m}\sqrt{p\Big(p+\frac{2\nu}{\nu-2}\Big)\times\frac{\log n}{n}} \;=\; \frac{\sigma_m}{C_m}\sqrt{p\times\frac{(\nu-2)p+2\nu}{\nu-2}\times\frac{\log n}{n}}\,,$$

which implies

$$\inf_{\epsilon>0}\Big(\varepsilon_{\nu;n,m,T}^{\mathrm{SGDw}}(\epsilon)\Big)^2 \;\le\; \Big(\varepsilon_{\nu;n,m,T}^{\mathrm{SGDw}}(3C_m\delta)\Big)^2$$
$$\le\; \frac{9\sigma_m^2 p((\nu-2)p+2\nu)\log n}{(\nu-2)n} \;+\; \kappa_{\nu;m}^2 2^{\frac{\nu-2}{\nu}}(r_\Psi)^{\frac{(\nu-2)p}{\nu}}(1+2\delta^{-1})^{\frac{(\nu-2)p}{\nu}}\exp\Big(-\frac{(\nu-2)p+2\nu}{2\nu}\log n\Big)$$
$$\le\; \frac{9\sigma_m^2 p((\nu-2)p+2\nu)\log n}{(\nu-2)n}$$
$$+\; \kappa_{\nu;m}^2 2^{\frac{\nu-2}{\nu}}(r_\Psi)^{\frac{(\nu-2)p}{\nu}}\Big(1+\frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2}((\nu-2)p+2\nu)^{1/2}}\frac{\sqrt{n}}{\sqrt{\log n}}\Big)^{\frac{(\nu-2)p}{\nu}} n^{-\frac{(\nu-2)p+2\nu}{2\nu}}$$
$$\le\; \frac{9\sigma_m^2 p((\nu-2)p+2\nu)\log n}{(\nu-2)n}$$
$$+\; \kappa_{\nu;m}^2 2^{\frac{\nu-2}{\nu}}(r_\Psi)^{\frac{(\nu-2)p}{\nu}}\Big(1+\frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2}((\nu-2)p+2\nu)^{1/2}}\Big)^{\frac{(\nu-2)p}{\nu}} n^{\frac{(\nu-2)p}{2\nu}-\frac{(\nu-2)p+2\nu}{2\nu}}$$
$$\le\; \Bigg(\frac{9\sigma_m^2 p((\nu-2)p+2\nu)}{\nu-2} + \kappa_{\nu;m}^2 2^{\frac{(\nu-2)}{\nu}}(r_\Psi)^{\frac{(\nu-2)p}{\nu}}\Big(1+\frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2}((\nu-2)p+2\nu)^{1/2}}\Big)^{\frac{(\nu-2)p}{\nu}}\Bigg)\frac{\log n}{n}\,.$$

Taking a squareroot across and using $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b>0$ gives the desired bound. The limiting result follows by substituting this bound into Theorem B.1. □

Proof of Lemma B.4. Let $\delta > 0$ and N_δ be defined as in the proof of Lemma B.3, with $N_\delta \le \big(r_\Psi(1+2/\delta)\big)^p$. Let $(\psi_l)_{l=1}^{N_\delta}$ be the centers of the covering δ-balls. The covering-ball argument of the proof of Lemma 3.1 of [21] shows that

$$\vartheta_{n,m,T}^{\mathrm{SGDw}}(3C_m\delta) \;\le \mathbb{P}\Big(\sup_{\psi\in\Psi}\Big\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\phi\big(K_\psi^m(X_i)\big)\big|X_i\big] - \mathbb{E}\big[\phi(K_\psi^m(X_i))\big]\Big\| > 3C_m\delta\Big)$$
$$\le \sum_{l=1}^{N_\delta}\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\phi\big(K_{\psi_l}^m(X_i)\big)\big|X_i\big] - \mathbb{E}\big[\phi(K_{\psi_l}^m(X_i))\big]\Big\| \ge C_m\delta\Big)\,.$$

By a Markov's inequality followed by the Burkholder's inequality applied to an average of i.i.d. summands (see e.g. [51] for $\nu > 2$), there exists a constant $C_\nu > 0$ depending only on ν such that

$$\vartheta_{n,m,T}^{\mathrm{SGDw}}(3C_m\delta) \;\le\; \sum_{l=1}^{N_\delta}\frac{\mathbb{E}\Big\|\sum_{i=1}^n \mathbb{E}\big[\phi\big(K_{\psi_l}^m(X_i)\big)\big|X_i\big] - \mathbb{E}\big[\phi(K_{\psi_l}^m(X_i))\big]\Big\|^\nu}{n^\nu C_m^\nu\delta^\nu}$$
$$\le \sum_{l=1}^{N_\delta}\frac{\mathbb{E}\big\|\mathbb{E}\big[\phi\big(K_{\psi_l}^m(X_1)\big)\big|X_i\big] - \mathbb{E}\big[\phi(K_{\psi_l}^m(X_1))\big]\big\|^\nu}{n^{\nu/2}C_m^\nu\delta^\nu}$$
$$\le \frac{N_\delta\kappa_{\nu;m}^\nu}{n^{\nu/2}C_m^\nu\delta^\nu} \;\le\; \frac{(r_\Psi)^p(1+2\delta^{-1})^p\kappa_{\nu;m}^\nu}{n^{\nu/2}C_m^\nu\delta^\nu}\,,$$

where we have used assumption A4 in the last line. This implies

$$\begin{aligned}\left(\varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(3C_m\delta)\right)^2 &= 9C_m^2\delta^2 + \kappa^2_{\nu;m}\left(\vartheta^{\mathrm{SGDw}}_{n,m,T}(3C_m\delta)\right)^{(\nu-2)/\nu} \\ &\leq 9C_m^2\delta^2 + \kappa^{\nu}_{\nu;m}(r_\Psi)^{\frac{(\nu-2)p}{\nu}} C_m^{-(\nu-2)} \times \frac{(1+2\delta^{-1})^{\frac{(\nu-2)p}{\nu}}}{n^{\frac{\nu-2}{2}}\delta^{\nu-2}} \ .\end{aligned}$$

Choosing $\delta = n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}} \leq 1$, we get that

$$\begin{aligned}\inf_{\epsilon>0}\left(\varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon)\right)^2 &\leq \left(\varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(3C_m\delta)\right)^2 \\ &\leq 9C_m^2 n^{-\frac{(\nu-2)\nu}{\nu^2+(\nu-2)p}} + \kappa^{\nu}_{\nu;m}(r_\Psi)^{\frac{(\nu-2)p}{\nu}} C_m^{-(\nu-2)} 3^{\frac{(\nu-2)p}{\nu}}\, n^{-\frac{(\nu-2)\nu}{\nu^2+(\nu-2)p}} \\ &= \left(9C_m^2 + \kappa^{\nu}_{\nu;m}(r_\Psi)^{\frac{(\nu-2)p}{\nu}} C_m^{-(\nu-2)} 3^{\frac{(\nu-2)p}{\nu}}\right) n^{-\frac{(\nu-2)\nu}{\nu^2+(\nu-2)p}} \ .\end{aligned}$$

Taking a squareroot across and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ gives the desired bound. The limiting result follows by substituting this bound into Theorem B.1. □

Proof of Lemma B.5. By a Markov's inequality and a Jensen's inequality with respect to the empirical average, we have

$$\begin{aligned}\vartheta^{\mathrm{SGDw}}_{n,m,T}(\epsilon) &\leq \sup_{t\in[T], j\in[N]} \frac{\sum_{i=1}^n \mathbb{E}\left\|\mathbb{E}\left[\phi\left(K^m_{\psi^{\mathrm{SGDw}}_{t-1,j}}(X_i)\right)\middle| X_i, \psi^{\mathrm{SGDw}}_{t-1,j}\right] - \mathbb{E}\left[\phi\left(K^m_{\psi^{\mathrm{SGDw}}_{t-1,j}}(X'_1)\right)\middle|\psi^{\mathrm{SGDw}}_{t-1,j}\right]\right\|}{n\epsilon} \\ &=: \sup_{t\in[T], j\in[N]} \frac{\sum_{i=1}^n A_{tji}}{n\epsilon} \ .\end{aligned}$$

Meanwhile by a triangle inequality and the ergodicity assumption, we have

$$\begin{aligned}A_{tji} &\leq \mathbb{E}\left\|\mathbb{E}\left[\phi\left(K^m_{\psi^{\mathrm{SGDw}}_{t-1,j}}(X_i)\right)\middle| X_i, \psi^{\mathrm{SGDw}}_{t-1,j}\right] - \mathbb{E}\left[\phi\left(X_1^{\psi^{\mathrm{SGDw}}_{t-1,j}}\right)\middle|\psi^{\mathrm{SGDw}}_{t-1,j}\right]\right\| \\ &\quad + \mathbb{E}\left\|\mathbb{E}\left[\phi\left(X_1^{\psi^{\mathrm{SGDw}}_{t-1,j}}\right)\middle|\psi^{\mathrm{SGDw}}_{t-1,j}\right] - \mathbb{E}\left[\phi\left(K^m_{\psi^{\mathrm{SGDw}}_{t-1,j}}(X'_1)\right)\middle|\psi^{\mathrm{SGDw}}_{t-1,j}\right]\right\| \\ &\leq 2\tilde{C}_K\tilde{\alpha}^m \ .\end{aligned}$$

This implies $\vartheta^{\mathrm{SGDw}}_{n,m,T}(\epsilon) \leq 2\tilde{C}_K\tilde{\alpha}^m\epsilon^{-1}$, and therefore

$$\left(\varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon)\right)^2 \leq \epsilon^2 + 2^{\frac{\nu-2}{\nu}}\kappa^2_{\nu;m}(\tilde{C}_K)^{\frac{\nu-2}{\nu}}\tilde{\alpha}^{\frac{(\nu-2)m}{\nu}}\epsilon^{-\frac{\nu-2}{\nu}} \ .$$

Choosing $\epsilon = \tilde{\alpha}^{(\nu-2)m/(3\nu-2)}$ gives

$$\inf_{\epsilon>0}\left(\varepsilon^{\mathrm{SGDw}}_{\nu;n,m,T}(\epsilon)\right)^2 \leq \left(1 + 2^{\frac{\nu-2}{\nu}}\kappa^2_{\nu;m}(\tilde{C}_K)^{\frac{\nu-2}{\nu}}\right)\tilde{\alpha}^{\frac{2(\nu-2)m}{3\nu-2}} \ .$$

Taking a squareroot across and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ gives the desired bound. The limiting result follows by substituting this bound into Theorem B.1. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract summarizes the theoretical results of our paper. We cleary mention that while CD *can* achieve these rates, assumptions are needed to guarantee this. We also mention the model class studied in this paper, namely (unnormalized) exponential family distributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention limitations multiple times in the main body of our work: after introducing the assumptions in Section 3.1, we mention that spectral gaps may not be verified for heavy tail distributions, or act numerically unfavorably for multimodal distributions, and that constants may be large in practice. In the discussion (Section 6), we reflect back on these limitations, and we mention the limitations of our considered model class (unnormalized exponential families).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All of our theoretical results are formally proved in the supplementary material. In the main paper, we provide some proof sketches (discussion of the tail term in Section 4, paragraphs after theorems in section 3) to provide high level intuitions behind the proofs. All of the theoretical results provided by our work are proved in the supplementary material, usually by invoking (and referencing) more atomic lemmas, also proved in the supplementary material. All of our theorems and lemmas are stated with the assumptions they require.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments, and thus no code or data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: As a theory paper, we believe that our work does not breach the code of conduct. With this work, we aim to advance the understanding of statistically efficient algorithms, a domain which has the potential to improve the overall *computational* efficiency of machine learning algorithms.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We believe that our work will not have the negative impacts mentioned (e.g., disinformation, surveillance, fairness, privacy, security considerations). We believe that advancing the domain of statistical effiency has the potential to improve the overall computational efficiency of machine learning algorithms, which, all else left equal, could constitute a positive societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Since our work does not include experiments nor data, and only analyzes existing models, we do not believe that our work requires any specific guidelines.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our work does not release any assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.