ANALYZING AND IMPROVING THE OPTIMIZATION LANDSCAPE OF NOISE-CONTRASTIVE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise-contrastive estimation (NCE) is a statistically consistent method for learning unnormalized probabilistic models. It has been empirically observed that the choice of the noise distribution is crucial for NCE's performance. However, such observations have never been made formal or quantitative. In fact, it is not even clear whether the difficulties arising from a poorly chosen noise distribution are statistical or algorithmic in nature. In this work, we formally pinpoint reasons for NCE's poor performance when an inappropriate noise distribution is used. Namely, we prove these challenges arise due to an ill-behaved (more precisely, flat) loss landscape. To address this, we introduce a variant of NCE called *eNCE* which uses an exponential loss and for which *normalized gradient descent* addresses the landscape issues *provably* when the target and noise distributions are in a given exponential family.

1 INTRODUCTION

Noise contrastive estimation (NCE) is a method for learning parameterized statistical models (Gutmann & Hyvärinen, 2010; 2012). To estimate a distribution P_*, NCE trains a discriminant model to distinguish between samples of P_* and a known distribution Q of our choice, often referred to as the "noise" distribution. If the function class for the discriminant model is representationally powerful enough, the optimal model learns the density ratio p_*/q, from which we can extract the density p_* since q is known (Menon & Ong, 2016; Sugiyama et al., 2012). Compared to the well-studied maximum likelihood estimation (MLE), NCE avoids calculating the (often intractable) partition function, [1] while maintaining the asymptotic consistency of MLE (Gutmann & Hyvärinen, 2012).

It is empirically well-documented that the choice of the noise distribution Q is crucial to both the statistical and algorithmic efficiency of NCE (Gutmann & Hyvärinen, 2010; 2012; Rhodes et al., 2020; Goodfellow et al., 2014; Gao et al., 2020). However, it has been observed in practice that even when following the standard guidelines for choosing Q, NCE can still yield parameter estimates far from the ground truth (Rhodes et al., 2020; Goodfellow et al., 2014; Gao et al., 2020). Most recently, Rhodes et al. (2020) identified a phenomenon they call the "density chasm," observing empirically that NCE performs poorly when the KL divergence between P_* and Q is large. One example is when P_*, Q are both tightly concentrated unimodal distributions with faraway modes; the region between the two modes will have a small density under both distributions, thus forming a "chasm". While it makes intuitive sense that NCE does not perform well under such settings—since disparate Q and P_* are easy to distinguish and do not require the model to learn much about P_* in order to do well on the classification task—there has not been a theoretical analysis of this phenomenon. In fact, it is not even clear whether the difficulty is statistical or algorithmic in nature.

In this work, we formally study the challenges for NCE with a fixed Q with a focus on distributions in an exponential family. We show that when the noise distribution Q is poorly chosen, the loss landscape can become extremely flat: in particular, even when P^* and Q are two univariate Gaussian with unit variance, the loss gradient and curvature can become exponentially small in the difference in their means. We prove that this poses challenges for standard first order and even second-order optimization methods, forcing them to take an exponential number of steps to converge to a good parameter estimate. Thus, standard approaches to minimizing convex functions such as

[1] The partition function is also known as the normalizing constant of an unnormalized density, such that the density after normalization will integrate to 1.

gradient descent—or even more advanced techniques such as momentum or Newton's method—are not suited to the NCE objective unless Q is close to P_* in KL sense.

To remedy this issue, we study an alternative method for optimizing the NCE objective. We consider instead Normalized Gradient Descent (NGD) whereby the gradient is normalized to have unit norm at each time step. Perhaps surprisingly, we prove that this small modification can overcome the problem of poor curvature in the Gaussian example. In general, we show the number of steps for NGD to converge to a good solution for the NCE loss depends on the *condition number* κ of the Hessian of the loss at the optimum—the growth of this condition number is unclear for P^* and Q when they belong to an exponential family.

To address this, we propose the *eNCE* loss, a variant to NCE that replaces the log loss in NCE with an exponential loss, and we show that the resulting condition number is polynomial in the dimension and the parameter distance between P^* and Q when they belong to an exponential family. Our proposed change of loss and optimization algorithm *together* form the first solution that provides a provable polynomial rate for learning the parameters of the ground truth distribution. Theoretically, both NCE and eNCE can potentially suffer from numerical issues during optimization when P^* and Q are far—this is an interesting direction for future work. Nonetheless, we find this to be a simple and effective fix to the flatness of the loss landscape in many settings, as evidenced by experimental results on synthetic and MNIST dataset.

1.1 Related Work

NCE and its variants have inspired a large volume of research in NLP (Mnih & Teh, 2012; Mnih & Kavukcuoglu, 2013; Dyer, 2014; Kong et al., 2020) as well as computer vision (Oord et al., 2018; Hjelm et al., 2018; Henaff, 2020; Tian et al., 2020). It has been observed empirically that NCE with a fixed noise Q is often insufficient for learning good generative models. The predominant class of approaches that have been proposed to overcome this issue aim to do so by not using a fixed Q but by iteratively solving multiple NCE problems with an updated Q, or equivalently updated discriminators. This includes the famous generative adversarial network (GAN) by Goodfellow et al. (2014), which uses a separate discriminator network updated throughout training. In a similar vein, Gao et al. (2020) also aimed to increase the discriminative power as the density estimator improves, and parameterize Q explicitly with a flow model. More recently, Rhodes et al. (2020) proposed the telescoping density ratio estimation, or TRE, which sidesteps the chasm by expanding p_*/q into a series of intermediate density ratios, each of which is easier to estimate, leading to strong empirical performance—though their work carries no formal guarantees.

With respect to a fixed Q, it remains an open question about what formally are the nature of the challenges posed by a poorly chosen Q, which could be statistical and/or algorithmic. Various previous works have analyzed the asymptotic behavior of NCE and its variants (Gutmann & Hyvärinen, 2012; Riou-Durand et al., 2018; Uehara et al., 2020), but these do not provide guidance on the finite step convergence of NCE or its common variants. The improvements to NCE in prior works are all borne out by the empirical observations of NCE practitioners, rather than motivated by theory, which is precisely the aim of this work.

Finally, we would like to note that prior work has proposed "generalized NCE" (Pihlaja et al., 2010; Gutmann & Hirayama, 2011; Uehara et al., 2020), which relates the NCE objective to minimizing the Bregman divergence. Generalized NCE says that we can design a family of training objectives by using different convex functions to define the Bregman divergence, and the proposed eNCE is an instance of the generalized NCE objective. The difference between these prior work and ours is again the different focuses on asymptotic behavior versus finite step convergence.

2 Preliminaries

The NCE objective Let P_* denote an unknown distribution in a parametric family $\{P_\theta\}_{\theta\in\Theta}$, for some bounded convex set Θ, with $P_* = P_{\theta_*}$. Our goal is to estimate P_* via P_θ for some $\theta \in \Theta$ by solving a noise contrastive estimation task. The noise distribution Q belongs to the same parametric family with parameters $\theta_q \in \Theta$, so that $Q = P_{\theta_q}$. We use p_θ, p_*, q to denote the probability density functions (pdfs) of P_θ, P_*, and Q; we may omit θ in P_θ, p_θ when it is clear from the context and write P, p instead. Given P_* and Q, the NCE loss of P is defined as follows:

Definition 2.1 (NCE Loss)**.** *The NCE loss of P_θ w.r.t. data distribution P_* and noise Q is:*

$$L(P_\theta) = -\frac{1}{2}\mathbb{E}_{P_*} \log \frac{p_\theta}{p_\theta + q} - \frac{1}{2}\mathbb{E}_Q \log \frac{q}{p_\theta + q} \tag{2.1}$$

The NCE loss can be interpreted as the binary cross-entropy loss for the classification task of distinguishing the data samples from the noise samples. Moreover, the NCE loss has a unique minimizer:

Lemma 2.1 (Gutmann & Hyvärinen 2012)**.** *The NCE objective in Definition 2.1 is uniquely minimized at $P = P_*$, provided that the support of Q covers that of P_*.*

Exponential family. We focus our attention on the exponential family, where the pdf for a distribution with parameter θ is $p_\theta(x) = \exp\left(\theta^\top \tilde{T}(x) - A(\theta)\right)$, with $\tilde{T}(x)$ denoting the sufficient statistics and $A(\theta)$ the log partition function. [2] The partition function is treated as a parameter in NCE, so we use τ to denote the extended parameter, i.e. $\tau := [\theta, \alpha]$ where α is the estimate for the log partition function. We accordingly extend the sufficient statistics as $T(x) = [\tilde{T}(x), -1]$ to account for the log partition function. The pdf with the extended representation is now simply $p_\tau(x) = \exp(\tau^\top T(x))$. We will use the notation P_θ and P_τ interchangeably. We will also use $\tau(\theta)$ to denote the log-partition extended parameterization when the log partition function α properly normalizes the distribution specified by θ.

A compelling reason for focusing on the exponential family is the observation that the NCE loss is convex in the parameter τ:

Lemma 2.2 (NCE convexity)**.** *For exponential family $p_{\theta,\alpha}(x) = h(x)\exp(\theta^\top \tilde{T}(x) - \alpha)$, the NCE loss is convex in parameter $\tau := [\theta, \alpha]$.*

Lemma 2.2 has been stated under more general settings by Uehara et al. (2020); an alternative self-contained proof is included in Appendix A for completeness.

Recall that Θ denotes the set of parameters without the extended coordinate for the log partition function. We assume the following on distributions supported on Θ:

Assumption 2.1 (Bounded parameter norm)**.** $\|\theta\|_2 \leq \omega$, $\forall \theta \in \Theta$.

Assumption 2.2 (Lipschitz log partition function)**.** *Assume the log partition function is β_Z-Lipschitz, that is, $\forall \theta_1, \theta_2 \in \Theta$, $|\log Z(\theta_1) - \log Z(\theta_2)| \leq \beta_Z \|\theta_1 - \theta_2\|$.*

Assumption 2.3 (Bounded singular values of the population Fisher matrix)**.** *There exist $\lambda_{\max}, \lambda_{\min} > 0$, such that $\forall \theta \in \Theta$, we have $\sigma_{\max}(\mathbb{E}_\theta[T(x)T(x)^\top]) \leq \lambda_{\max}$, and $\sigma_{\min}(\mathbb{E}_\theta[T(x)T(x)^\top]) \geq \lambda_{\min}$.*

Assumption 2.4 (Smooth change in the Fisher matrix)**.** *Assume the maximum and minimum singular values of the Fisher matrix change smoothly. Namely, there exist constants $\gamma_{\max}, \gamma_{\min} > 0$ s.t.*

$$\|\nabla_\theta \sigma_{\max}(\mathbb{E}_\theta[T(x)T(x)^\top])\| \leq \gamma_{\max}, \ \|\nabla_\theta \sigma_{\min}(\mathbb{E}_\theta[T(x)T(x)^\top])\| \leq \gamma_{\min}$$

Assumptions 2.2-2.4 can be viewed as smoothness assumptions on the first, second and third order derivatives of the log partition function, and can in fact be viewed as introducing *structural parameters* of the distributions, rather than extra constraints. For example, distributions with flatter tails will have a larger $\lambda_{\max}$, which then translates to a slower rate in the results; distributions that are closer to being singular with have a smaller $\lambda_{\min}$, etc. In particular, Assumption 2.3 says the singular values of the Fisher matrix $\mathbb{E}_\theta[T(x)T(x)^\top]$ should be bounded from above and below. It can be shown that the Fisher matrix is proportional to the Hessian of the NCE objective when using $Q = P_*$, which means Assumption 2.3 can be interpreted as saying the NCE task can be solved efficiently under the optimal choice of Q.

3 Overview of results

We first provide an informal overview of our results, focusing on learning of exponential families.

[2] Another common format of the exponential family PDF is $p_\theta(x) = h(x)\exp\left(\theta^\top T(x) - A(\theta)\right)$ where $h(x)$ is a non-negative function. Such $h(x)$ could be absorbed into $\tilde{T}(x)$ and θ with corresponding coordinates $\log(h(x))$ and 1.

Flatness of population landscape: Our first contribution is a negative result identifying a key source of difficulty for NCE optimization to be an ill-behaved *population* landscape. We show that due to an extremely flat landscape, gradient descent or Newton's method with *standard choices* of step sizes will need to take an exponential number of steps to find a reasonable parameter estimate.

We emphasize that though Gaussian mean estimation is a trivial task, its simplicity *strengthens the results above*: we are proving a *negative* result so that failures with a simpler setup means a stronger result. Moreover, the results only apply to *standard* choices of step sizes, such as inversely proportional to the smoothness for gradient descent, or to the ratio between the smoothness and strong convexity for Newton's method. This does not rule out the possibility that a cleverly designed learning rate schedule or a different algorithm would work efficiently; the results are however still meaningful since gradient descent with standard step sizes is the most common choice in practice.

Overcoming flatness using normalized gradient descent: Our second contribution is to show that the flatness problem can be solved by a simple modification to gradient descent if the loss is well-conditioned. Specifically, we show that the convergence rate for *normalized gradient descent* is polynomial in the parameter distance and κ_*, the *condition number* of the Hessian at the optimum. One immediate consequence is that in Gaussian mean estimation, for a target error of $\delta \in (0, \frac{1}{\beta_Z}]$ in parameter distance, NCE optimized with NGD achieves a rate of $O(\frac{1}{\delta^2})$, which is the same as the optimal rate achieved by MLE.

The remaining question is then whether κ_* is polynomial in the parameters of interests. We show that κ_* can be related to the Bhattacharyya coefficient between P_* and Q, which indeed grows polynomially in parameter distance under certain assumptions as detailed in Section 5.2.

Polynomial condition number for the eNCE loss: Our third and final contribution is that if we modify the NCE objective slightly—namely, use the exponential loss in place of the log loss—then the condition number at the optimum is guaranteed to be polynomial. We call this new objective *eNCE* . Combined with the NGD result, we get that running NGD on the eNCE objective achieves a polynomial convergence guarantee.

We then provide empirical evidence on synthetic and MNIST dataset that eNCE with NGD performs comparatively with NGD on the original NCE loss, and both outperform gradient descent.

4 FLATNESS OF THE NCE LOSS

In this section, we study the challenges posed to NCE when using a badly chosen fixed Q. The main thrust of the results is to show that both algorithmic and statistical challenges can arise because the NCE loss is *poorly behaved*, particularly for first- and second-order optimization algorithms: when P_*, Q are far, the loss landscape is extremely flat near the optimum. In particular, the gradient has exponentially small norm and the strong convexity constant decreases exponentially fast, limiting the convergence rate of the excess risk. We further show that when moving from $P = Q$ to $P = P_*$, the loss drops from $\Theta(1)$ to a value that is exponentially small in terms of the distance between P_* and Q. Consequently, common gradient-based and second order methods will take exponential number of steps to converge.

An important note is that our analysis is at the population level, implying that the hardness comes from the landscape itself regardless of the statistical estimators used.

Setup – Gaussian mean estimation: For the negative results in this section, let's consider an exceedingly simple scenario of 1-dimensional, fixed-variance Gaussian mean estimation. We will demonstrate the enormous difficulty of achieving a good parameter estimate, even for such a simple problem—this bodes ill for NCE objectives corresponding to more complex models in practice, which certainly pose a much more difficult challenge. In particular, let P_*, Q, P be Gaussians with identity variance. Let $\theta_*, \theta_q, \theta$ denote the respective means, with θ_* being the target mean that NCE aims to estimate. When the covariance is known to be 1, we can denote $h(x) := \exp(-\frac{x^2}{2})$, and parametrize the pdf of a 1d Gaussian with mean θ as $p(x) = h(x)\exp(\langle \tau(\theta), T(x)\rangle)$,[3] where the parameter is $\tau(\theta) := [\theta, \frac{\theta^2}{2} + \log\sqrt{2\pi}]$ and the sufficient statistics are $T(x) := [x, -1]$. [4] We will shorthand $\tau(\theta)$ when it is clear from the context.

[3]Thus, we are setting h to be the base measure for the exponential family we are considering.

[4]Recall that the last coordinate -1 acts as a sufficient statistic for the log partition function.

Without loss of generality, we will assume $\theta_q = 0$, and $\theta_* > 0$, and denote $R := \theta_* - \theta_q$. We will write $\tau_* := \tau(\theta_*) = [R, \frac{R^2}{2} + \log\sqrt{2\pi}]$, and $\tau_q := \tau(\theta_q) = [0, \log\sqrt{2\pi}]$. As a clarification, the results stated in this section will be in terms of R, hence the asymptotic notations Ω, O never hide dominating dependency on R. [5]

4.1 Properties of the NCE loss

We first describe several properties of the NCE loss that will be useful in the analysis of first- and second-order algorithms.

To start, we show that the dynamic range of the loss is large: that is, the optimal NCE loss is exponentially small as a function of R; on the other hand, if θ is initialized close to θ_q, the initial loss would be on the order of a constant. Precisely:

Proposition 4.1 (Range of NCE loss)**.** *Consider the 1d Gaussian mean estimation task with mean* $\theta_*, \theta_q \in \mathbb{R}$*, and a known variance of 1. Denote* $R := |\theta_q - \theta_*|$ *where* $R \gg 1$*, Then, the loss at* $\theta = \theta_q$ *is* $\log 2$*, while the minimal loss* L_* *is* $L_*(R) = c\exp(-R^2/8)$ *for some* $c \in [\frac{1}{2}, 2]$.

The next shows we *need to* decrease the loss to be on an order comparable to the optimum value. Namely, the loss is very flat close to θ_*, thus in order to recover a θ close to θ_*, we have to reach a very small value for the loss. Precisely:

Proposition 4.2. *Under the same setup as Proposition 4.1, for a given* $\delta \in (0, 1)$*, if the learned parameter* τ *satisfies* $\|\tau - \tau^*\|_2 \leq \delta$*, then* $L(\tau) - L(\tau^*) \leq R\exp(-R^2/8)\,\delta^2$.

The way we will leverage Propositions 4.1 and 4.2 to prove lower bounds is to say that if the updates of an iterative algorithm are too small, the convergence will take an exponential number of steps.

Proposition 4.2 is proven via the Taylor expansion at θ^*: since the gradient is 0 at θ_*, we just need to bound the Hessian at θ_*. We show:

Lemma 4.1 (Smoothness at $P = P^*$)**.** *Under the same setup as Proposition 4.1, the smoothness at* $P = P_*$ *is upper bounded as* $\sigma_{\max}(\nabla^2 L(\tau_*)) \leq \frac{R}{\sqrt{2\pi}}\exp(-R^2/8)$.

We will also need a bound on the strong convexity constant (i.e. smallest singular value) at $P = P^*$:

Lemma 4.2 (Strong convexity at $P = P^*$)**.** *Under the same setup as Proposition 4.1, the minimum singular value at* $P = P_*$ *is* $\sigma^*_{\min}(\nabla^2 L(\tau_*)) = \Theta\left(\frac{1}{R}\exp\left(-\frac{R^2}{8}\right)\right)$.

Finally, in order to estimate the choice of the step size for standard optimization methods, we will also need a bound of the smoothness at $P = Q$:

Lemma 4.3 (Smoothness at $P = Q$)**.** *Under the same setup as Proposition 4.1, the smoothness at* $P = Q$ *is lower bounded as* $\sigma_{\max}(\nabla^2 L(\tau_q)) \geq \frac{R^2}{2}$.

Lemma 4.1, 4.2 are proved in Appendix D.4, and Lemma 4.3 is proved in Appendix D.5.

4.2 Lower bounds on first- and second-order methods

With the landscape properties at hand, we are now ready to provide lower bounds for both first-order and second-order methods. For first-order methods, we show that:

Theorem 4.1 (Lower bound for gradient-based methods)**.** *Let* P_*, Q, P *be 1d Gaussian with variance 1. Assume* $\theta_q = 0, \theta_* > 0$ *without loss of generality, and assume* $R := \theta_* - \theta_q \gg 1$*. Then, gradient descent with any step size* $\eta = o(1)$ *from an initialization* $\tau = \tau_q$ *will need an exponential number of steps to reach some* τ' *that is* $O(1)$ *close to* τ_*.

Note, the maximum step size $\eta = o(1)$ the theorem applies to is actually a loose bound: the standard setting of step size for gradient descent is $\eta \leq 1/\lambda_M$ for $\lambda_M := \max_{\theta\in\Theta}\sigma_{\max}(\nabla^2 L(\tau(\theta)))$, which is $\Omega(R^2)$ by Lemma 4.3. Theorem 4.1 helps explain why NCE with a far-away Q fails in practice, if we set the budget for the number of updates to be polynomial.

[5] For example, for $R \gg 1$, $R\exp(R^2) = O(\exp(R^2))$, but the constant in $O(1)$ will not depend on R.

A natural remedy to the drastically changing norms of the gradients is to use methods that can properly precondition the gradient. This motivates the use of second order methods, which adapt to the geometry of the loss and hence can potentially perform more competitively.

Unfortunately, standard second-order approaches are again of no help, and the number of steps required to converge remains exponential. Consider Newton's method with updates of the form $\eta(\nabla^2 L)^{-1}\nabla L$. At first glance, this looks like it may solve the issue of a flat gradient, since the Hessian $\nabla^2 L$ may also be exponentially small hence canceling out with the exponentially small gradient. However, the flatness of the landscape forces us to take an exponentially small step size η, resulting in the following claim:

Theorem 4.2 (Lower bound for Newton's method)**.** *Let P_*, Q, P satisfy the same conditions as in theorem 4.1. Let $\lambda_\rho := \min_{\theta\in\Theta} \sigma_{\min}(\nabla^2 L(\tau_\theta))$, $\lambda_M := \max_{\theta\in\Theta} \sigma_{\max}(\nabla^2 L(\tau_\theta))$. Then, running the Newton's method with step size $\eta = O(\frac{\lambda_\rho}{\lambda_M})$ from an initialization $\tau = \tau_q$ will need an exponential number of steps to reach some τ' that is $O(1)$ close to τ_*.*

Again, the condition $\eta = O\left(\frac{\lambda_\rho}{\lambda_M}\right)$ follows the typical step size choice for Newton's method, i.e. the step size should be upper bounded by the ratio between the *global* strong convexity constant and the *global* smoothness of the function, which is exponentially small for this setup by Lemma 4.2, 4.3.

5 Normalized gradient descent for well-conditioned losses

We have seen that due to an ill-behaved landscape, NCE optimized with standard gradient descent or Newton's method will fail to reach a good parameter estimate efficiently, even on a problem as simple as Gaussian mean estimation, and even with access to the population gradient.

In this section, we will show that a close relative of gradient descent, *normalized gradient descent* (NGD), despite its simplicity, provides a fix to the flatness problem to exponential family distributions when the Hessian of the loss is well-conditioned close to the optimum.

Precisely, recall that the NGD updates for a loss function L is $\tau_{t+1} = \tau_t - \eta\frac{\nabla L(\tau_t)}{\|\nabla L(\tau_t)\|_2}$. We assume that in a neighborhood around τ_*, the change in the shape of the Hessian $\boldsymbol{H}$ is moderate: [6]

Assumption 5.1 (Hessian in a neighborhood of τ_*)**.** *Under assumption 2.2 with constant β_Z, assume that for any τ such that $\|\tau - \tau_*\|_2 \leq \frac{1}{\beta_Z}$, it holds that $\sigma_{\max}(\boldsymbol{H}(\tau)) \leq \beta_u \cdot \sigma_{\max}(\boldsymbol{H}(\tau_*))$, and $\sigma_{\min}(\boldsymbol{H}(\tau)) \geq \beta_l \cdot \sigma_{\min}(\boldsymbol{H}(\tau_*))$, for some constant $\beta_u, \beta_l > 0$.*

The main result of this section states that NGD can find a parameter estimate efficiently for exponential families, where the number of steps required is polynomial in the distance between the initial estimate and the optimum:

Theorem 5.1. *Let L be any loss function that is convex in the exponential family parameter and satisfies Assumptions 5.1 and 2.1 - 2.4. Furthermore, let P_*, Q be exponential family distributions with parameters τ_*, τ_q and let κ_* be the condition number of the Hessian at $P = P_*$. Then, for any $0 < \delta \leq \frac{1}{\beta_Z}$ and parameter initialization τ_0, with step size $\eta \leq \sqrt{\frac{\beta_l}{\beta_u \kappa_*}}\delta$, performing NGD on the population objective L guarantees that after $T \leq \frac{\beta_u \kappa_*}{\beta_l} \cdot \frac{\|\tau_0 - \tau_*\|^2}{\delta^2}$ steps, there exists an iterate $t \leq T$ such that $\|\tau_t - \tau_*\|_2 \leq \delta$.*

The main technical ingredient for proving Theorem 5.1 is the following Lemma:

Lemma 5.1. *Suppose Assumptions 2.2 and 5.1 hold with constants β_Z, β_u and β_l. Let L be a convex function with minimizer τ_*, and let $g := \nabla L(\tau)$. For any $\delta \leq \frac{1}{\beta_Z}$, let $\gamma = \sqrt{\frac{\beta_l}{\beta_u \kappa_*}}\delta$. Then for all τ s.t. $\|\tau - \tau_*\|_2 \geq \delta$, we have $L(\tau_* + \gamma\frac{g}{\|g\|}) \leq L(\tau)$.*

Lemma 5.1 explains the dependency on κ_* in the NGD convergence rate. The intuition of the proof is that in a small neighborhood around τ_*, the set of parameters that have the same loss form an "ellipsoid", and by Taylor expansion, any two points in the same set will have a distance-to-τ_* ratio upper bounded roughly by $\sqrt{\kappa_*}$. The details are in Appendix C.2.

[6] As a concrete example, we will show in the next section that a variant of NCE satisfies both conditions.

Proof sketch for Theorem 5.1: the proof leverages two observations: the convexity of NCE loss on exponential family parameters, and that the Hessian in a neighborhood around τ_* changes moderately as stated in Assumption 5.1. One can then show there exists a global constant γ such that for any τ_t satisfying $\|\tau_t - \tau_*\|_2 \geq \delta$, one step of NGD update guarantees a decrease of γ^2 in the squared error, which means NGD must have found an δ-close estimate within $\frac{\|\tau_0 - \tau_*\|^2}{\gamma^2}$ steps. The full proof is deferred to Appendix C.1.

5.1 Example: 1d Gaussian mean estimation

It is relatively straightforward to check that NGD addresses the flatness problem faced by Gaussian mean estimation we considered in Section 4:

Corollary 5.1. *Let P_*, Q be 1d Gaussian with covariance 1 and mean $\theta_* = R$ where $R \ll 1$, and $\theta_q = 0$. For any given $\delta \leq \frac{1}{R}$ and initial estimate $\tau_0 = \tau_q$, NGD can find an estimate τ such that $\|\tau - \tau_*\|_2 \leq \delta$, with at most $O(\frac{R^6}{\delta^2})$ steps.*

Intuitively, the effectiveness of NGD comes from the crucial observation that though the magnitude for the loss and derivatives can be exponentially small, they share the same exponential factor, making normalization effective. Formally, it can be shown that $\frac{\beta_u}{\beta_l} = O(1)$ (Appendix D.6). Corollary 5.1 then follows from Theorem 5.1 and the curvature and strong convexity from Lemma 4.1, 4.2.

5.2 Bounds on the condition number of NCE

The convergence rate in Theorem 5.1 depends on κ_*, the condition number of the NCE Hessian at the optimum, and Hessian-related constants β_u, β_l in Assumption 5.1. We now show that under the setup of Theorem 5.1, κ_* and β_u, β_l can be related to the *Bhattacharyya coefficient* between P_* and Q, which is a similarity measure defined as $\text{BC}(P_*, Q) := \int_x \sqrt{p_*(x)q(x)}dx$. As a result, we get the following convergence guarantee:

Theorem 5.2. *Suppose Assumptions 2.1, 2.3 hold with constants ω, $\lambda_{\max}$, and $\lambda_{\min}$. Consider a NCE task with data distribution P_1 and noise distribution P_2, parameterized by $\theta_1, \theta_2 \in \Theta$ respectively. Then for any given $\delta \leq \frac{1}{R}$ and initial estimate $\tau_0 = \tau_q$, NGD finds an estimate τ such that $\|\tau - \tau_*\|_2 \leq \delta$ within $T \leq C \cdot \frac{1}{BC(P_*,Q)^3} \frac{\|\tau_0 - \tau_*\|^2}{\delta^2}$ steps, where $C := 18 \exp\left(\frac{2}{\beta_Z}\right) \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^3 \cdot \min\left\{\frac{2\lambda_{\max}^2}{\lambda_{\min}^2}, \frac{2\lambda_{\min} + \gamma_{\max}\|\bar{\delta}\|}{\lambda_{\min} - \gamma_{\min}\|\bar{\delta}\|}\right\}$.*

In particular, when P_*, Q are not too far, we can further show a lower bound on $\text{BC}(P_*, Q)$:

Lemma 5.2. *For P_1, P_2 parameterized by $\theta_1, \theta_2 \in \Theta$, if $\|\theta_1 - \theta_2\|_2^2 \leq \frac{4}{\lambda_{\max}}$, then $BC(P_1, P_2) \geq \frac{1}{2}$.*

The proofs of Theorem 5.2 and Lemma 5.2 rely on analyzing the geodesic on the manifold of square root densities $\sqrt{p}$ equipped with the Hellinger distance as a metric; the details are deferred to Appendix C.3 and C.4. It is also worth noting that Theorem 5.2 only requires $\|\theta_1 - \theta_2\|$ to be smaller than a constant, rather than tending to zero as usually required for analyses using Taylor expansions.

Finally, we would like to note that although our analysis can be tightened, it is unlikely to remove such dependency since NGD only uses first-order information. [7] Moreover, the condition number κ_* also affects the practical use of Newton-like methods, since matrix inversion is widely known to be sensitive to numerical issues when the matrix is extremely ill-conditioned. It is an interesting open question whether a non-standard preconditioning approach might be amenable to this setting.

6 Analyzing eNCE : NCE with an exponential loss

The previous section proved that NGD can serve as a simple fix to overcome the flatness problem of NCE for well-conditioned losses. However, though we showed κ_* has a polynomial growth when the distributions P, Q^* are sufficiently close —it is unclear how κ_* behaves beyond this threshold.

[7] In the next section, we will that the condition number is provably polynomial in $\|\theta_* - \theta_q\|$ for a variant of the NCE loss.

In this section, we introduce a slight modification to the NCE objective, which we call the *eNCE* objective, for which κ_* depends *polynomially* on some class-related constants. This means though eNCE may still suffer from the flatness problem, *eNCE and NGD together provide a solution that guarantees a polynomial convergence rate*.

Towards formalizing this, the eNCE loss is defined as:

Definition 6.1 (eNCE Loss)**.** *Let* $\varphi(x) := \log\sqrt{\frac{p(x)}{q(x)}}$*, and* $l(x,y) := \exp(-y\varphi(x))$ *for* $y \in \{\pm 1\}$*. The eNCE loss of* P_θ *w.r.t. data distribution* P_* *and noise* Q *is:*

$$L_{\exp}(P_\theta) = \frac{1}{2}\mathbb{E}_{x\sim P_*}[l(x,1)] + \frac{1}{2}\mathbb{E}_{x\sim P_*}[l(x,-1)] = \frac{1}{2}\int_x p_*\sqrt{\frac{q(x)}{p(x)}} + \frac{1}{2}\int_x q\sqrt{\frac{p(x)}{q(x)}} \tag{6.1}$$

It can be checked easily that the minimizing φ learns $\varphi(x) = \frac{1}{2}\log\frac{p_*}{q}$. Moreover, each φ is associated with an induced distribution p, defined by $p(x) = \exp(\varphi(x))q(x)$.

Relation to NCE: Same as the original NCE loss (referred to as "NCE" below), eNCE learns to solve a distinguishing task between samples from P_* or Q. The difference lies only in the losses, which have analogous forms: the NCE loss described in Def. 2.1 can be rewritten in the same form with $l(x,y) := \log\frac{1}{1+\exp(-y\psi(x))}$ and $\psi(x) := \log\frac{p(x)}{q(x)}$.

The main advantage of the exponential loss is that the Hessian at the optimum is now guaranteed to be well-conditioned:

Lemma 6.1 (Polynomial condition number for eNCE loss)**.** *Under Assumption 2.3 with constants* $\lambda_{\max}, \lambda_{\min}$*, the condition number of the eNCE Hessian at the optimum is bounded by* $\kappa_* \le \frac{\lambda_{\max}}{\lambda_{\min}}$*.*

We can also show that eNCE satisfies part (ii) of Assumption 5.1. Due to space considerations, the proof is deferred to Appendix B.1.

Lemma 6.2. *Under assumption 2.2 with constant* β_Z*, for any unit vector* $\boldsymbol{u}$ *and constant* $c \in [0, \frac{1}{\beta_Z}]$*, the maximum and minimum singular values of* $\boldsymbol{H}(\tau_* + c\boldsymbol{u})$ *satisfy assumption 5.1 with constants* $\beta_u = 2\exp(1)\cdot\frac{\lambda_{\max}}{\lambda_{\min}}$*,* $\beta_l = \frac{1}{2\exp(1)}\cdot\frac{\lambda_{\min}}{\lambda_{\max}}$*.*

Lemma 6.1 and Lemma 6.2 together imply the Hessian is well-conditioned around the optimum. Combined with Theorem 5.1, we have the main result of this section:

Theorem 6.1. *Let* P_*, Q *be exponential family distributions with parameters* τ_*, τ_q *under Assumption 2.1-2.4. Let* $\lambda_{\max}, \lambda_{\min}$ *be constants for Assumption 2.3. For any given* $\delta \le \frac{1}{\beta_Z}$ *and parameter initialization* τ_0*, performing NGD on the eNCE objective guarantees that when taking* $T \le 4\exp(2)\cdot\frac{\lambda_{\max}^3}{\lambda_{\min}^3}\cdot\frac{\|\tau_0-\tau_*\|^2}{\delta^2}$ *steps, there exists an iterate* $t \le T$ *such that* $\|\tau_t - \tau_*\|_2 \le \delta$*.*

Proof. Theorem 6.1 follows directly from Theorem 5.1, using the condition number bound from Lemma 6.1 and constants from 6.2. □

6.1 Proof of Lemma 6.1

It can be checked that the Hessian at $P = P_*$ is $\boldsymbol{H}_* := \nabla^2 L(P_*) = \frac{1}{4}\int_x \sqrt{p_* q}\nabla\log p(\nabla\log p)^\top$. Recall that $\theta_*, \theta_q, \tilde{T}$ denote the parameters and sufficient statistics without the partition function coordinate, and τ_*, τ_q, T denote the extended version with the partition function, e.g. $\tau_* = [\theta_*, \log Z(\theta_*)]$, $T(x) = [\tilde{T}(x), -1]$. Then, we can rewrite $\boldsymbol{H}_*$ as:

$$\begin{aligned}
\boldsymbol{H}_* =& \frac{1}{4}\int_x \sqrt{p_* q}T(x)T(x)^\top = \frac{1}{4}\int_x \exp\left(\frac{(\tau_*+\tau_q)^\top}{2}T(x)\right)T(x)T(x)^\top \\
=& \frac{1}{4}\int_x \exp\left(\frac{(\theta_*+\theta_q)^\top}{2}\tilde{T}(x) - \frac{1}{2}\log Z(\theta_*) - \frac{1}{2}\log Z(\theta_q)\right)T(x)T(x)^\top \\
=& \frac{1}{4}\underbrace{\frac{Z\left(\frac{\theta_*+\theta_q}{2}\right)}{\sqrt{Z(\theta_*)Z(\theta_q)}}}_{B(P_*,Q)}\int_x \frac{\exp\left((\frac{\theta_*+\theta_q}{2})^\top\tilde{T}(x)\right)}{Z\left(\frac{\theta_*+\theta_q}{2}\right)}T(x)T(x)^\top dx = \frac{B(P_*,Q)}{4}\mathbb{E}_{\frac{\theta_*+\theta_q}{2}}[TT^\top]
\end{aligned} \tag{6.2}$$

The Lemma then follows from $\lambda_{\min}\boldsymbol{I} \preceq \mathbb{E}_{\frac{\theta_*+\theta_q}{2}}[TT^\top] \preceq \lambda_{\max}\boldsymbol{I}$ by Assumption 2.3.

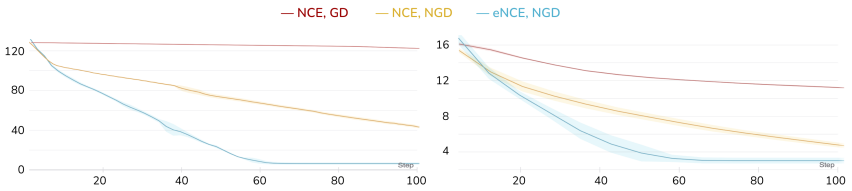

Figure 1: Results for estimating 1d (left) and 16d (right) Gaussians, plotting $\min_{t\in[T]}\|\tau_* - \tau_t\|_2$ (y-axis) against the number of updates T (x-axis). In both cases, when using NCE, normalized gradient descent ("NCE, NGD", yellow) largely outperforms gradient descent ("NCE, GD", red). When using NGD, the proposed eNCE ("eNCE, NGD", blue) decays faster than the original NCE loss. The results are averaged over 5 runs, with shaded areas showing the standard deviation.

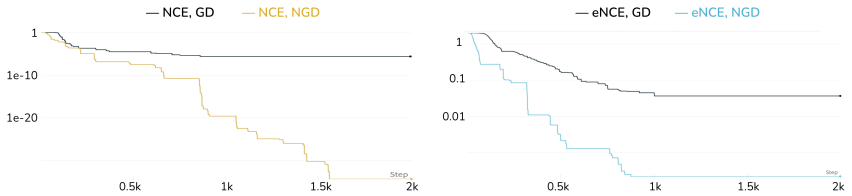

Figure 2: Results on MNIST, plotting loss value (y-axis, log scale) against update steps (x-axis). The left plot shows NCE optimized by GD (black) and NGD (yellow), and the right shows eNCE optimized by GD (black) and NGD (blue). It can be seen that NGD outperforms GD in both cases.

7 Empirical verification

To corroborate our theory, we verify the effectiveness of NGD and eNCE on Gaussian mean estimation and the MNIST dataset. For MNIST, we use a ResNet-18 to model the log density ratio $\log(p/q)$, following the setup in TRE (Rhodes et al., 2020).

Results: For Gaussian data, we run gradient descent (GD) and normalized gradient descent (NGD) on the NCE loss and eNCE loss. Figure 1 compares the best runs under each setup given a fixed computation budget (100 update steps), where "best" is defined to be the run with the lowest loss on fresh samples. The plots show the minimum parameter distance $\min_{t\in[T]}\|\tau_* - \tau_t\|_2$ for each step T. We find that NGD indeed outperforms GD, and that the proposed eNCE sees a further improvement over NCE while additionally enjoying provable polynomial convergence guarantees. For MNIST, we can no longer compare parameter distances since τ_* is unknown. Instead, we compare the result of optimization directly in terms of loss achieved, again under a fixed computation budget (2K steps). The results are shown in Figure 2, with NGD converging significantly faster for both NCE and eNCE. We note that eNCE can be numerically unstable, especially when P_*, Q are well separated. Implementation details to prevent numerical issues are included in Appendix E.

8 Conclusion and Discussions

We provided a theoretical analysis of the algorithmic difficulties that arise when optimizing the NCE objective with an uninformative noise distribution, stemming from an ill-behaved loss landscape. Our theoretical results are inspired by empirical observations in prior works (Rhodes et al., 2020; Gao et al., 2020; Goodfellow et al., 2014) and provide the first formal explanation on the nature of the optimization problems of NCE. Our negative results showed that even on the simple task of Gaussian mean estimation, and even assuming access to the population gradient, gradient descent and Newton's method with standard step size choice still require an exponential number of steps to reach a good solution.

We then proposed modifications to the NCE loss and optimization algorithm, whose combination results in the first provably polynomial convergence rate for NCE. The loss we propose, eNCE, can be efficiently optimized using normalized gradient descent and empirically outperforms existing methods. We hope these theoretical results will help identify promising new directions in the search for simple, effective, and practical improvements to noise-contrastive estimation.

REFERENCES

Jacob Andreas, Maxim Rabinovich, Dan Klein, and Michael I Jordan. On the accuracy of self-normalized log-linear models. *arXiv preprint arXiv:1506.04147*, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Accurate and conservative estimates of mrf log-likelihood using reverse annealing. In *Artificial Intelligence and Statistics*, pp. 102–110. PMLR, 2015.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

Chris Dyer. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*, 2014.

Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7518–7528, 2020.

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

Charles J Geyer. On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):261–274, 1994.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Roger B Grosse, Chris J Maddison, and Russ R Salakhutdinov. Annealing between distributions by averaging moments. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/fb60d411a5c5b72b2e7d3527cfc84fd0-Paper.pdf.

Michael Gutmann and Jun-ichiro Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

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*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13 (2), 2012.

Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.

Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pp. 197–209. PMLR, 2014.

Elad Hazan, Kfir Y Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *arXiv preprint arXiv:1507.02030*, 2015.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *arXiv preprint arXiv:1605.06336*, 2016.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Syx79eBKwr`.

Matthieu Labeau and Alexandre Allauzen. Learning with noise-contrastive estimation: Easing training by learning to scale. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3090–3101, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1261`.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pp. 304–313. PMLR, 2016.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, 26:2265–2273, 2013.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1751–1758, 2012.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Miika Pihlaja, Michael Gutmann, and Aapo Hyvärinen. A family of computationally efficient and simple estimators for unnormalized statistical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *arXiv preprint arXiv:2006.12204*, 2020.

Lionel Riou-Durand, Nicolas Chopin, et al. Noise contrastive estimation: Asymptotic properties, formal comparison with mc-mle. *Electronic Journal of Statistics*, 12(2):3473–3518, 2018.

Akash Srivastava, Kai Xu, Michael U. Gutmann, and Charles Sutton. Generative ratio matching networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJg7spEYDS`.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, USA, 1st edition, 2012. ISBN 0521190177.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.

Che-Ping Tsai, Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. Heavy-tailed streaming statistical estimation. *arXiv preprint arXiv:2108.11483*, 2021.

Masatoshi Uehara, Takafumi Kanamori, Takashi Takenouchi, and Takeru Matsuda. A unified statistically efficient estimation framework for unnormalized models. In *International Conference on Artificial Intelligence and Statistics*, pp. 809–819. PMLR, 2020.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. In *International Conference on Machine Learning*, pp. 10957–10967. PMLR, 2020.