
The Convergence of Contrastive Divergences

Alan Yuille

Department of Statistics
University of California at Los Angeles
yuille@stat.ucla.edu

1 Introduction

Section (2) describes CD and shows that it is closely related to a class of stochastic approximation algorithms for which convergence results exist. In section (3) we state and give a proof of a simple convergence theorem for stochastic approximation algorithms. Section (4) applies the theorem to give sufficient conditions for convergence of CD.

2 CD and its Relations

the model distribution: $P(x|\omega) = e^{-E(x;\omega)} / Z(\omega)$. to try performing steepest descent

on $D(P_0(x)||P(x|\omega))$. The *steepest descent algorithm*:

$$\omega_{t+1} - \omega_t = \gamma_t \left\{ - \sum_x P_0(x) \frac{\partial E(x;\omega)}{\partial \omega} + \sum_x P(x|\omega) \frac{\partial E(x;\omega)}{\partial \omega} \right\}, \quad (1)$$

where the $\{\gamma_t\}$ are constants.

CD:

$$\omega_{t+1} - \omega_t = \gamma_t \left\{ - \sum_x P_0(x) \frac{\partial E(x;\omega)}{\partial \omega} + \sum_x Q_\omega(x) \frac{\partial E(x;\omega)}{\partial \omega} \right\},$$

where $Q_\omega(x)$ is the empirical distribution function on the samples obtained by initializing the chain at the data samples $P_0(x)$ and running the Markov chain forward for m steps.

We now observe that CD is similar to a class of *stochastic approximation algorithms* which also use MCMC methods to stochastically approximate the second term on the rhs of eq (1). These algorithms are reviewed in [7] and have been used, for example, to learn probability distributions for modelling image texture [8].

(2)

This SAA, and its many variants, have been extensively studied and convergence results have been obtained [7]. The convergence results are based on stochastic approximation theorems [6] whose history starts with the analysis of the Robbins-Monro algorithm [5]. Precise conditions can be specified which guarantee convergence in probability. In particular, Kushner [9] has proven convergence to global optima. Within the NIPS community, Orr and Leen [10] have studied the ability of these algorithms to escape from local minima by basin hopping.

3 Stochastic Approximation Algorithms and Convergence

The general SAA:

$$\omega_{t+1} = \omega_t - \gamma_t S(\omega_t, N_t), \quad (3)$$

where N_t is a rv sampled from a distribution $P_n(N)$, γ_t is the damping coefficient, and $S(.,.)$ is an arbitrary function.

We now state a theorem which gives sufficient conditions to ensure that the SAA (3) converges to a (solution) state ω^* . The theorem is chosen because of the simplicity of its proof and we point out that a large variety of alternative results are available [6,7,9]. The theorem involves three basic concepts:

1. a function $L(\omega) = (1/2)|\omega - \omega^*|^2$ a measure of the distance of the current state ω from the solution state ω^* (in the next section we will require $\omega^* = \arg \min_{\omega} D(P_0(x) || P(x|\omega))$)
2. the expected value $\sum_N P_n(N) S(\omega, N)$ of the update term in the SAA(3)
3. the expected squared magnitude $\langle |S(\omega, N)|^2 \rangle$ of the update term.

The theorem states that the algorithm will converge provided three conditions are satisfied. These conditions are fairly intuitive. The first condition requires that the expected update $\sum_N P_n(N) S(\omega, N)$ has a large component towards the solution ω^* (i.e. in the direction of the negative gradient of $L(\omega)$). The second condition requires that the expected squared magnitude $\langle |S(\omega, N)|^2 \rangle$ is bounded, so that the “noise” in the update is not too large. The third condition requires that the damping coefficients γ_t decrease with time t , so that the algorithm eventually settles down into a fixed state. This condition is satisfied by setting $\gamma_t = 1/t$, $\forall t$ (which is the fastest fall off rate consistent with the SAC theorem).

Stochastic Approximation Convergence (SAC) Theorem . let $L(\omega) = (1/2)|\omega - \omega^*|^2$.

Then SAA (3) $\rightarrow \omega^*$ with probability 1 provided:

- (1) $-\nabla L(\omega) \cdot \sum_N P_n(N) S(\omega, N) \geq C L(\omega)$,
- (2) $\langle |S(\omega, N)|^2 \rangle_t \leq C (1 + L(\omega))$, where the expectation $\langle . \rangle_t$ is taken wrt all the data prior to time t
- (3) $\sum_{t=1}^{\infty} \gamma_t = \infty$ and $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$.

Proof. The proof [12] is a consequence of the supermartingale convergence theorem [11]. This theorem states that if X_t, Y_t, Z_t are positive random variables obeying $\sum_{t=0}^{\infty} Y_t \leq \infty$ with probability one and $\langle X_{t+1} \rangle \leq X_t + Y_t - Z_t$, $\forall t$, then X_t converges with probability 1

and $\sum_{t=0}^{\infty} Z_t < \infty$. To apply the theorem, set $X_t = (1/2)|\omega_t - \omega^|^2$, set $Y_t = (1/2)K_2\gamma_t^2$ and $Z_t = -X_t(K_2\gamma_t^2 - K_1\gamma_t)$ (Z_t is positive for sufficiently large t). Conditions 1 and 2 imply that X_t can only converge to 0. The result follows after some algebra.*

4 CD and SAC

The CD algorithm can be expressed as SAA :

$$S(\omega_t, N_t) = - \sum_x P_0(x) \frac{\partial E(x; \omega)}{\partial \omega} + \sum_x Q_{\omega}(x) \frac{\partial E(x; \omega)}{\partial \omega}, \quad (4)$$

where N_t corresponds to the MCMC sampling used to obtain $Q_{\omega}(x)$.

We conjecture that weaker conditions, such as requiring only that the gradient of $E(x; \omega)$ be bounded by a function linear in ω , can be obtained using the more sophisticated martingale analysis described in [7].

to understand the first condition and to determine whether the solution is unbiased. These require studying the *expected CD update*:

$$\sum_{N_t} P_n(N_t) S(\omega_t, N_t) = - \sum_x P_0(x) \frac{\partial E(x; \omega)}{\partial \omega} + \sum_{y, x} P_0(y) K_\omega^m(y, x) \frac{\partial E(x; \omega)}{\partial \omega}, \quad (5)$$

which is derived using the fact that the expected value of $Q_\omega(x)$ is $\sum_y P_0(y) K_\omega^m(y, x)$ (running the transition kernel m times).

We now re-express this expected CD update in two different ways, Results 1 and 2, which give alternative ways of understanding it. We then proceed to Results 3 and 4 which give conditions for convergence and unbiasedness of CD.

Background material from Markov Chain theory [13]

We choose the transition kernel $K_\omega(x, y)$ to satisfy *detailed balance* so that $P(x|\omega)K_\omega(x, y) = P(y|\omega)K_\omega(y, x)$. Detailed balance is obeyed by many MCMC algorithms and, in particular, is always satisfied by Metropolis-Hasting algorithms. It implies that $P(x|\omega)$ is the invariant kernel of $K_\omega(x, y)$ so that $\sum_x P(x|\omega)K_\omega(x, y) = P(y|\omega)$ (all transition kernels satisfy $\sum_y K_\omega(x, y) = 1, \forall x$).

Detailed balance implies that the matrix $Q_\omega(x, y) = P(x|\omega)^{1/2} K_\omega(x, y) P(y|\omega)^{-1/2}$ is symmetric and hence has orthogonal eigenvectors and eigenvalues $\{e_\omega^\mu(x), \lambda_\omega^\mu\}$. The eigenvalues are ordered by magnitude (largest to smallest). The first eigenvalue is $\lambda^1 = 1$ (so $|\lambda^\mu| < 1, \mu \geq 2$). By standard linear algebra, we can write $Q_\omega(x, y)$ in terms of its eigenvectors and eigenvalues $Q_\omega(x, y) = \sum_\mu \lambda_\omega^\mu e_\omega^\mu(x) e_\omega^\mu(y)$, which implies that we can express the transition kernel applied m times by:

$$K_\omega^m(x, y) = \sum_\mu \{\lambda_\omega^\mu\}^m \{P(x|\omega)\}^{-1/2} e_\omega^\mu(x) \{P(y|\omega)\}^{1/2} e_\omega^\mu(y) = \sum_\mu \{\lambda_\omega^\mu\}^m u_\omega^\mu(x) v_\omega^\mu(y), \quad (6)$$

where the $\{v_\omega^\mu(x)\}$ and $\{u_\omega^\mu(x)\}$ are the *left and right eigenvectors* of $K_\omega(x, y)$:

$$\begin{aligned} \mathbf{K} &= \mathbf{V} \mathbf{D} \mathbf{U}^T = \mathbf{p} \mathbf{o} \mathbf{1} + \mathbf{V} \mathbf{D} \mathbf{1} \mathbf{U}^T \\ \mathbf{V}^T \mathbf{U} &= \mathbf{U}^T \mathbf{V} = \mathbf{1} \end{aligned}$$

$$v_\omega^\mu(x) = e_\omega^\mu(x) \{P(x|\omega)\}^{1/2}, \quad u_\omega^\mu(x) = e_\omega^\mu(x) \{P(x|\omega)\}^{-1/2}, \quad \forall \mu, \quad (7)$$

and it can be verified that $\sum_x v_\omega^\mu(x) K_\omega(x, y) = \lambda_\omega^\mu v_\omega^\mu(y)$, $\forall \mu$ and $\sum_y K_\omega(x, y) u_\omega^\mu(y) = \lambda_\omega^\mu u_\omega^\mu(x)$, $\forall \mu$. In addition, the left and right eigenvectors are ~~mutually orthonormal~~ so that

$$\sum_x v_\omega^\mu(x) u_\omega^\nu(x) = \delta_{\mu\nu}. \quad \text{bi-orthnormal}$$

Moreover, the first left and right eigenvectors can be calculated explicitly:

$$v_\omega^1(x) = P(x|\omega), \quad u_\omega^1(x) \propto 1, \quad \lambda_\omega^1 = 1, \quad (9)$$

Result 1. The expected CD update corresponds to replacing the update term $\sum_x P(x|\omega) \frac{\partial E(x;\omega)}{\partial \omega}$ in the steepest descent equation (1) by:

$$\sum_x \frac{\partial E(x;\omega)}{\partial \omega} P(x|\omega) + \sum_{\mu=2} \{\lambda_{\omega}^{\mu}\}^n \left\{ \sum_y P_0(y) u_{\omega}^{\mu}(y) \right\} \left\{ \sum_x v_{\omega}^{\mu}(x) \frac{\partial E(x;\omega)}{\partial \omega} \right\}, \quad (10)$$

where $\{v_{\omega}^{\mu}(x), u_{\omega}^{\mu}(x)\}$ are the left and right eigenvectors of $K_{\omega}(x, y)$ with eigenvalues $\{\lambda^{\mu}\}$.

Result 1 demonstrates that the expected update of CD is similar to the steepest descent rule, see equations (1,10), but with an additional term $\sum_{\mu=2} \{\lambda_{\omega}^{\mu}\}^n \left\{ \sum_y P_0(y) u_{\omega}^{\mu}(y) \right\} \left\{ \sum_x v_{\omega}^{\mu}(x) \frac{\partial E(x;\omega)}{\partial \omega} \right\}$ which will be small provided the magnitudes of the eigenvalues $\{\lambda_{\omega}^{\mu}\}$ are small for $\mu \geq 2$ (or if the transition kernel can be chosen so that $\sum_y P_0(y) u_{\omega}^{\mu}$ is small for $\mu \geq 2$).

$$g := E' - P_{\infty} \cdot E'$$

Result 2. Let $g(x; \omega) = \frac{\partial E(x;\omega)}{\partial \omega} - \sum_x P(x|\omega) \frac{\partial E(x;\omega)}{\partial \omega}$, then $\sum_x P(x|\omega) g(x; \omega) = 0$, the extrema of the Kullback-Leibler divergence occur when $\sum_x P_0(x) g(x; \omega) = 0$, and the expected update rule can be written as:

$$\omega_{t+1} = \omega_t - \gamma_t \left\{ \sum_x P_0(x) g(x; \omega) - \sum_{y,x} P_0(y) K_{\omega}^m(y, x) g(x; \omega) \right\}. \quad (11)$$

1. $P_{\infty} \cdot g = 0$
2. w: KL-extrema iff $P_0 \cdot g = 0$
3. w: CD-fixed pt iff $P_0 \cdot g = P_0^T K^m g$

Result 3. The fixed points ω^* of the CD algorithm are true (unbiased) extrema of the KL divergence (i.e. $\sum_x P_0(x) g(x; \omega^*) = 0$) if, and only if, we also have $\sum_{y,x} P_0(y) K_{\omega^*}^m(y, x) g(x; \omega^*) = 0$. A sufficient condition is that $P_0(y)$ and $g(x; \omega)$ lie in orthogonal eigenspaces of $K_{\omega^*}(y, x)$. This includes the (known) special case when there exists ω^* such that $P(x|\omega^*) = P_0(x)$ [2].

w: CD-f.p. = KL-extrema (unbias)
iff $P_0 \cdot g = P_0^T K^m g = 0$

Proof. The second part can be obtained by the eigenspace analysis in Result 1.

Suppose $P_0(x) = P(x|\omega^*)$, $p_0 = p_{\infty}$

$v_{\omega^*}^1(x) = P(x|\omega^*)$, and so $\sum_y P_0(y) u_{\omega^*}^{\mu}(y) = 0$, $\mu \neq 1$. Moreover, $\sum_x v_{\omega^*}^1(x) g(x; \omega^*) = 0$

Result 3 shows that whether CD converges to an unbiased estimate usually depends on the specific form of the MCMC transition matrix $K_\omega(y, x)$. But there is an intuitive argument why the bias term $\sum_{y,x} P_0(y) K_{\omega^*}^m(y, x) g(x; \omega^*)$ may tend to be small at places where $\sum_x P_0(x) g(x; \omega^*) = 0$. This is because for small m , $\sum_y P_0(y) K_{\omega^*}^m(y, x) \approx P_0(x)$ which satisfies $\sum_x P_0(x) g(x; \omega^*) = 0$. Moreover, for large m , $\sum_y P_0(y) K_{\omega^*}^m(y, x) \approx P(x|\omega^*)$ and we also have $\sum_x P(x|\omega^*) g(x; \omega^*) = 0$.

Alternatively, using Result 1, the bias term $\sum_{y,x} P_0(y) K_{\omega^*}^m(y, x) g(x; \omega^*)$ can be expressed as $\sum_{\mu=2} \{\lambda_{\omega^*}^\mu\}^m \{\sum_y P_0(y) u_{\omega^*}^\mu(y)\} \{\sum_x v_{\omega^*}^\mu(x) \frac{\partial E(x; \omega^*)}{\partial \omega}\}$. This will tend to be small provided the eigenvalue moduli $|\lambda_{\omega^*}^\mu|$ are small for $\mu \geq 2$ (i.e. the standard conditions for a well defined Markov Chain). In general the bias term should decrease exponentially as $|\lambda_{\omega^*}^\mu|^m$. Clearly it is also desirable to define the transition kernels $K_\omega(x, y)$ so that the right eigenvectors $\{u_\omega^\mu(y) : \mu \geq 2\}$ are as orthogonal as possible to the observed data $P_0(y)$.

The practicality of CD depends on whether we can find an MCMC sampler such that the bias term $\sum_{y,x} P_0(y) K_{\omega^*}^m(y, x) g(x; \omega^*)$ is small for most ω . If not, then the alternative stochastic algorithms may be preferable.

Finally we give convergence conditions for the CD algorithm.

Result 4 *CD will converge with probability 1 to state ω^* provided $\gamma_t = 1/t$, $\frac{\partial E}{\partial \omega}$ is bounded, and*

$$(\omega - \omega^*) \cdot \left\{ \sum_x P_0(x) g(x; \omega) - \sum_{y,x} P_0(y) K_\omega^m(y, x) g(x; \omega) \right\} \geq C |\omega - \omega^*|^2, \quad (12)$$

Proof. This follows from the SAC theorem and Result 2. The boundedness of $\frac{\partial E}{\partial \omega}$ is required to ensure that the “update noise” is bounded in order to satisfy the second condition of the SAC theorem.

Results 3 and 4 can be combined to ensure that CD converges (with probability 1) to the correct (unbiased) solution. This requires specifying that ω^* in Result 4 also satisfies the conditions $\sum_x P_0(x) g(x; \omega^*) = 0$ and $\sum_{y,x} P_0(y) K_{\omega^*}^m(y, x) g(x; \omega^*) = 0$.