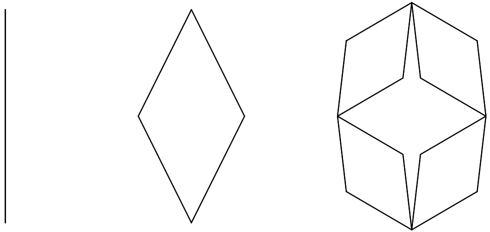
Normal Approximation for Hierarchical Sequences

Larry Goldstein

http://math.usc.edu/∼larry

The Diamond Lattice: Three Scales



Properties (e.g. conductance) at level 2 depend on properties at level 1, . . .

$$X_2 = F(\mathbf{X}_1), \qquad X_1 = F(\mathbf{X}_0)\ldots$$

Hierarchical Models

$$X_{n+1} = F(\mathbf{X}_n) \quad where \quad \mathbf{X}_n = (X_{n,1}, \ldots, X_{n,k})^\mathsf{T}$$

with $X_{n,i}$ independent, each with distribution X_n.

Conditions on F, due to Shneiberg, Li and Rogers, and Wehr, imply the weak law (here assumed)

$$X_n \to_p c,$$

and by Woo and Wehr which imply

$$W_n \to_d \mathcal{N}(0,1), \quad \text{for} \quad W_n = \frac{X_n - EX_n}{\sqrt{\mathsf{Var}(X_n)}}.$$

Classical Central Limit Theorem as Hierarchical Model

Taking F to give the *average*

$$F(x_1, x_2) = \frac{x_1 + x_2}{2}$$

gives in distribution

$$X_n = \frac{X_{0,1} + \cdots + X_{0,2^n}}{2^n}.$$

At stage n there are $N = 2^n$ variables, would expect a bound to the normal Z of the form

$$d(W_n, Z) \le C\gamma^n \quad \text{where} \quad \gamma^n = N^{-1/2} = (1/\sqrt{2})^n.$$

Averaging Functions

We say F is (strictly) averaging

1. $\min_i x_i \le F(\mathbf{x}) \le \max_i x_i$, and strictly when $\min_i x_i < \max_i x_i$.
2. $F(\mathbf{x}) \le F(\mathbf{y})$ whenever $x_i \le y_i$, and strictly when $x_j < y_j$ for some j.

Say F is scaled averaging when $F(\mathbf{x})/F(\mathbf{1}_k)$ is averaging.

Diamond Lattice Conductivity Function

Parallel and series resistor combination rules

$$L_1(x_1, x_2) = x_1 + x_2, \quad L_{-1}(x_1, x_2) = (x_1^{-1} + x_2^{-1})^{-1}$$

gives the weighted $w_i > 0$ diamond lattice conductivity function

$$F(\mathbf{x}) = \left(\frac{1}{w_1 x_1} + \frac{1}{w_2 x_2}\right)^{-1} + \left(\frac{1}{w_3 x_3} + \frac{1}{w_4 x_4}\right)^{-1},$$

a scaled averaging function.

Approximate Linear Recursion

Approximate $X_{n+1} = F(\mathbf{X}_n)$ by a linear recursion around the mean $c_n = EX_n$ with small perturbation R_n,

$$X_{n+1} = \boldsymbol{\alpha}_n \cdot \mathbf{X}_n + R_n, \quad n \geq 0,$$

where $\boldsymbol{\alpha}_n = F'(\mathbf{c}_n)$, $\mathbf{c}_n = (c_n, \ldots, c_n)^{\mathsf{T}} \in \mathbb{R}^k$, and F' the gradient of F.

Assuming the gradient $\boldsymbol{\alpha} = F'(\mathbf{c})$ at limit c is not a scalar multiple of a unit vector rules out trivial cases such as $F(x_1, x_2) = x_1$ and makes $\lambda = ||\boldsymbol{\alpha}|| < 1$.

Zero Bias Transformation

Goldstein and Reinert 1997: For W a mean zero variance σ^2 random variable, there exists W^* such that for all absolutely continuous f for which expectation $E|Wf(W)|$ exists,

$$EWf(W) = \sigma^2 Ef'(W^*).$$

From Stein's equation,

$$EZf(Z) = \sigma^2 Ef'(Z) \quad \text{if and only if} \quad Z \sim \mathcal{N}(0, \sigma^2).$$

Hence:

$$W^* =_d W \quad \text{if and only if} \quad W \sim \mathcal{N}(0, \sigma^2).$$

Zero and Size Biasing

For $W \geq 0$ with $EW < \infty$, we say W^s has the W-size bias distribution if for all f for which $E|Wf(W)| < \infty$,

$$EWf(W) = EWEf(W^s).$$

To zero (size) bias a sum

$$W = \sum_{i=1}^{k} X_i$$

of mean zero (non-negative) independent variables, pick one proportional to its variance (mean) and replace with biased version. See Goldstein and Rinott, 1996.

Wasserstein distance d

With $\mathcal{L} = \{g : \mathbb{R} \to \mathbb{R} : |g(y) - g(x)| \le |y - x|\}$ and,

$\mathcal{F} = \{f : f$ absolutely continuous, $f(0) = f'(0) = 0, f' \in \mathcal{L}\}$,

$$d(Y, X) = \sup_{g \in \mathcal{L}} |E(g(Y) - g(X))| = \sup_{f \in \mathcal{F}} |E(f'(Y) - f'(X))|.$$

Dual form, minimal L_1 distance

$$d(Y, X) = \inf E|Y - X|.$$

Zero Bias and distance d

Lemma 1 *Let W be a mean zero, finite variance random variable, and let W^* have the W-zero bias distribution. Then with d the Wasserstein distance, and Z a normal variable with the same variance as W,*

$$d(W,Z) \leq 2d(W,W^*).$$

Take $\sigma^2 = 1$. For $||h'|| \leq 1$, $||f''|| \leq 2$,

$$\begin{aligned} |Eh(W) - Nh| &= |E[f'(W) - Wf(W)]| \\ &= |[Ef'(W) - Ef'(W^*)]| \\ &\leq ||f''||E|W - W^*| \\ &\leq 2d(W,W^*). \end{aligned}$$

Zero Bias and distance δ

The L_1 distance $E|W^* - W|$ gives a bound on the Wasserstein distance between W^* and W, and therefore between W and Z, for Lipschitz functions (smooth). For $EW = 0$, $EW^2 = 1$, we can bound δ

$$\delta = \sup_{-\infty < x < \infty} |P(W \leq x) - P(Z \leq x)| \leq \gamma\,(86 + 12\gamma)$$

by a quantity which depends on the distribution of W only through the L_2 distance

$$\gamma^2 = E(W^* - W)^2.$$

Contraction Mapping in d

Lemma 2 *For* $\boldsymbol{\alpha} \in \mathbb{R}^k$ *with* $\lambda = ||\boldsymbol{\alpha}|| \neq 0$, *let*

$$Y = \sum_{i=1}^{k} \frac{\alpha_i}{\lambda} W_i,$$

where W_i *are mean zero, variance one, independent random variables distributed as* W*. Then*

$$d(Y, Y^*) \leq \varphi\, d(W, W^*),$$

and $\varphi = \sum_i |\alpha_i|^3 / (\sum_i \alpha_i^2)^{3/2} < 1$ *if and only if* $\boldsymbol{\alpha}$ *is not a scalar multiple of a unit vector.*

Contraction:Proof

With $\quad P(I=i)=\dfrac{\alpha_i^2}{\lambda^2}, \quad |Y-Y^*|=\dfrac{|\alpha_I|}{\lambda}|W_I-W_I^*|.$

Since $W_i =_d W$, we may take $(W_i, W_i^*) =_d (W, W^*)$

$$E|Y-Y^*|=\sum_{i=1}^{k}\frac{|\alpha_i|^3}{\lambda^3}E|W_i-W_i^*|=\varphi E|W-W^*|.$$

Choosing the pair W, W^* to achieve the infimum, we obtain

$$d(Y,Y^*)\le E|Y-Y^*|=\varphi E|W-W^*|=\varphi\, d(W,W^*).$$

Pause: The Classical CLT and d

Take W_i iid mean zero variance σ^2 and

$$Y = n^{-1/2} \sum_{i=1}^{n} W_i.$$

Setting $\alpha_i = n^{-1/2}$ gives $\varphi = n^{-1/2}$, and

$$d(Y, Z) \leq 2d(Y, Y^*) \leq 2n^{-1/2} d(W, W^*) \rightarrow 0$$

as $n \rightarrow \infty$, proof of the CLT with a bound in d and constant depending on $E|W^* - W| = ||W^* - W||_1$.

Pause (Protracted): The Classical CLT and δ

For any Y with $EY = 0, EY^2 = 1$ and $\gamma^2 = E(Y^* - Y)^2$,

$$\sup_{-\infty<x<\infty} |P(Y \le x) - P(Z \le x)| \le \gamma\,(86 + 12\gamma)\,.$$

Hence, for

$$Y = n^{-1/2}\sum_{i=1}^{n} W_i, \quad \text{we have} \quad Y^* - Y = n^{-1/2}(W_i^* - W_i).$$

In this case,

$$\gamma = n^{-1/2}\sqrt{E(W^* - W)^2} \quad \text{when} \quad W_i =_d W,$$

a proof of the CLT with bound in δ and constant depending on $||W^* - W||_2$.

Linear Iteration

Normalizing $X_{n+1} = \boldsymbol{\alpha}_n \cdot \mathbf{X}_n$, with $\lambda_n = ||\boldsymbol{\alpha}_n||$ and $\sigma_n^2 = \mathrm{Var}(X_n)$ we have

$$W_{n+1} = \sum_{i=1}^{k} \frac{\alpha_{n,i}}{\lambda_n} W_{n,i} \quad \text{with} \quad W_n = \frac{X_n - c_n}{\sigma_n}.$$

Iterated contraction gives

$$d(W_n, Z) \leq 2d(W_n, W_n^*) \leq 2 \left(\prod_{i=0}^{n-1} \varphi_i \right) d(W_0, W_0^*).$$

Non-linear Iteration

Let $X_{n+1} = \boldsymbol{\alpha}_n \cdot \mathbf{X}_n + R_n$, where $\mathbf{X}_n$ is a vector of iid variables distributed as X_n, $EX_n = c_n$, $\mathrm{Var}(X_n) = \sigma_n^2$, and $\lambda_n = ||\boldsymbol{\alpha}_n|| \neq 0$. Set

$$Y_n = \sum_{i=1}^{k} \frac{\alpha_{n,i}}{\lambda_n} W_{n,i} \quad \text{where} \quad W_n = \frac{X_n - c_n}{\sigma_n}$$

and (to measure the discrepancy from linearity)

$$\beta_n = E|W_{n+1} - Y_n| + \frac{1}{2} E|W_{n+1}^3 - Y_n^3|.$$

Theorem 1 *If there exist* $(\beta, \varphi) \in (0,1)^2$ *such that*

$$\limsup_{n\to\infty} \frac{\beta_n}{\beta^n} < \infty \quad \textit{and} \quad \limsup_{n\to\infty} \varphi_n = \varphi,$$

then with $\gamma = \beta$ *when* $\varphi < \beta$*, and for any* $\gamma \in (\varphi, 1)$ *when* $\beta \leq \varphi$*, there exists* C *such that*

$$d(W_n, Z) \leq C\gamma^n.$$

Apply Theorem 1 to hierarchical X_n.

Glimpse at Proof of Theorem 1

With $f' - wf = h - Nh$, $f \in \mathcal{F}$ implies

$$|h'(w)| \le (1 + 3w^2/2).$$

Usual role of h and f reversed:

$$\begin{aligned}
& |Ef'(W_{n+1}) - Ef'(W^*_{n+1})| \\
= \; & |Eh(W_{n+1}) - Nh| \\
= \; & |E\left(h(W_{n+1}) - h(Y_n) + h(Y_n) - Nh\right)| \\
\le \; & \beta_n + |Eh(Y_n) - Nh| \\
\le \; & \beta_n + |E(f'(Y_n) - f'(Y_n^*))| \\
\le \; & \beta_n + d(Y_n, Y_n^*) \\
\le \; & \beta_n + \varphi_n d(W_n, W_n^*).
\end{aligned}$$

Theorem 2 *Let* X_0 *be a non constant random variable with* $P(X_0 \in [a,b]) = 1$ *and* $X_{n+1} = F(\mathbf{X}_n)$ *with* $F : [a,b]^k \to [a,b]$*, twice continuously differentiable. Suppose* F *is averaging and that* $X_n \to_p c$*, with* $\boldsymbol{\alpha} = F'(\mathbf{c})$ *not a scalar multiple of a unit vector. Then with* Z *a standard normal variable, for all* $\gamma \in (\varphi, 1)$ *there exists* C *such that*

$$d(W_n, Z) \le C\gamma^n \quad \text{where} \quad \varphi = \frac{\sum_{i=1}^k |\alpha_i|^3}{(\sum_{i=1}^k |\alpha_i|^2)^{3/2}},$$

is a positive number strictly less than 1. The value φ *achieves a minimum of* $1/\sqrt{k}$ *if and only if the components of* $\boldsymbol{\alpha}$ *are equal.*

Averaging Networks

$(V, \mathcal{E}, \mathbf{w})$ is a weighted network with vertex set V, edge set $\mathcal{E}$ and non-negative weights $\mathbf{w}$, if V has two distinguished vertices, the *source* a and the *sink* b, $\mathcal{E} \subset (V \times V) \setminus \bigcup_{v \in V} (v, v)$. Without loss of generality the graph $(V, \mathcal{E})$ is connected and $w_i \in (0, \infty)$.

Theorem 3 *Let $(V, \mathcal{E}, \mathbf{w})$ be a weighted network with effective weight between any two components determined by scaled averaging and homogeneous parallel and series combination rules $P(x_1, x_2)$ and $S(x_1, x_2)$. Then the effective weight $F(\mathbf{x})$ between source and sink is a scaled averaging function.*

Special Case: Resistor Networks

Woo and Wehr show the conductance function of a resistor network, with the series and parallel combination rules,

$$L_1(x_1, x_2) = x_1 + x_2 \quad \text{and} \quad L_{-1}(x_1, x_2) = (x_1^{-1} + x_2^{-1})^{-1}$$

is scaled averaging. This is a special case of the above result for the weighted L_p norm functions, which are scaled averaging and homogeneous.

Fast Rates for the Diamond Lattice

Define the 'side equally weighted network' to be the one with $\mathbf{w} = (w, w, 2-w, 2-w)^\mathsf{T}$ for $w \in (1,2)$; such weights are positive and satisfy $F(\mathbf{w}) = 1$.

For $w = 1$ all weights are equal, and we have $\boldsymbol{\alpha} = 4^{-1}\mathbf{1}_4$, and hence φ achieves its minimum value $1/2 = 1/\sqrt{k}$ corresponding to the rate $N^{-1/2+\epsilon}$.

For $1 \le w < 2$ we have $1/2 \le \varphi < 1/\sqrt{2}$, the case $w \uparrow 2$ corresponding to the least favorable rate for the side equally weighted network of $N^{-1/4+\epsilon}$.

Slow Rates for the Diamond Lattice

With only the restriction that the weights are positive and satisfy $F(\mathbf{w}) = 1$ consider for $t > 0$,

$$\mathbf{w} = (1 + 1/t, s, t, 1/t)^{\mathsf{T}} \quad \text{where}$$

$$s = [(1 - (1/t + t)^{-1})^{-1} - (1 + 1/t)^{-1}]^{-1}.$$

When $t = 1$ we have $s = 1$ and $\varphi = 11\sqrt{2}/27$.

As $t \to \infty$, $s/t \to 1/2$ and $\boldsymbol{\alpha}$ tends to the unit vector $(1, 0, 0, 0)$, so $\varphi \to 1$.

Since $11\sqrt{2}/27 < 1/\sqrt{2}$, the diamond lattice rate can achieve any γ in the range $(1/2, 1)$, corresponding to $N^{-\theta}$ for any $\theta \in (0, 1/2)$.