1 Gradient descent, natural gradient descent, mirror gradient descent, and curved families

We work in the ambient exponential family

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\big(\langle \boldsymbol{\eta}, \boldsymbol{T}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta})\big), \qquad \boldsymbol{\eta} \in \mathcal{N} \subset \mathbb{R}^d,$$

with dual coordinates

$$\boldsymbol{\eta}^* = \boldsymbol{\nabla} A(\boldsymbol{\eta}) \in \mathcal{N}^*, \qquad \boldsymbol{\eta} = \boldsymbol{\nabla} A^*(\boldsymbol{\eta}^*).$$

For data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, define

$$\bar{\boldsymbol{T}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{T}(\boldsymbol{x}_n),$$

and consider the negative log-likelihood

$$L(\boldsymbol{\eta}) = A(\boldsymbol{\eta}) - \langle \boldsymbol{\eta}, \bar{\boldsymbol{T}} \rangle.$$

Its gradient is

$$\boldsymbol{\nabla} L(\boldsymbol{\eta}) = \boldsymbol{\eta}^* - \bar{\boldsymbol{T}},$$

so the optimum is characterized by moment matching,

$$\boldsymbol{\eta}^* = \bar{\boldsymbol{T}}.$$

1.1 Gradient descent

As usual, take an exponential family

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\big(\langle \boldsymbol{\eta}, \boldsymbol{T}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta})\big), \qquad \boldsymbol{\eta} \in \mathcal{N} \subset \mathbb{R}^d,$$

with dual coordinates

$$\boldsymbol{\eta}^* = \boldsymbol{\nabla} A(\boldsymbol{\eta}) \in \mathcal{N}^*, \qquad \boldsymbol{\eta} = \boldsymbol{\nabla} A^*(\boldsymbol{\eta}^*).$$

For data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, define

$$\bar{\boldsymbol{T}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{T}(\boldsymbol{x}_n),$$

and consider the negative log-likelihood

$$L(\boldsymbol{\eta}) = A(\boldsymbol{\eta}) - \langle \boldsymbol{\eta}, \bar{\boldsymbol{T}} \rangle, \qquad \boldsymbol{\nabla}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}) = \boldsymbol{\eta}^* - \bar{\boldsymbol{T}},$$

so the optimum is characterized by the usual moment matching

$$\boldsymbol{\eta}^* = \bar{\boldsymbol{T}}.$$

Ordinary gradient descent (GD) updates $\boldsymbol{\eta}$ by subtracting the gradient in the chosen coordinates:

$$\boldsymbol{\eta}^i_{t+1} = \boldsymbol{\eta}^i_t - \alpha_t \, \delta^{ij} \partial_j L(\boldsymbol{\eta}_t).$$

Equivalently,

$$\boldsymbol{\eta}_{t+1} = \arg\min_{\boldsymbol{\eta}} \left\{ \langle \boldsymbol{\nabla}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}_t), \boldsymbol{\eta} - \boldsymbol{\eta}_t \rangle + \frac{1}{2\alpha_t} \|\boldsymbol{\eta} - \boldsymbol{\eta}_t\|^2 \right\}.$$

This uses the Euclidean metric δ_{ij} as an external choice. In index language, $\partial_j L$ is naturally a covector, and the update inserts δ^{ij} to force it into a vector. Ordinary GD is therefore coordinate-dependent.

For the exponential-family likelihood this becomes

$$\boldsymbol{\eta}^i_{t+1} = \boldsymbol{\eta}^i_t - \alpha_t \, \delta^{ij} (\boldsymbol{\eta}^*_{t,j} - \bar{T}_j).$$

It does not reflect the underlying geometry of the family.

1.2 Natural gradient descent

The exponential family carries the Fisher metric

$$g_{ij}(\boldsymbol{\eta}) = \partial_i \partial_j A(\boldsymbol{\eta}), \qquad g^{ij}(\boldsymbol{\eta}^*) = \partial^{*i} \partial^{*j} A^*(\boldsymbol{\eta}^*).$$

Natural gradient descent (NGD) replaces the Euclidean metric by this intrinsic one:

$$\boldsymbol{\eta}^i_{t+1} = \boldsymbol{\eta}^i_t - \alpha_t \, g^{ij}(\boldsymbol{\eta}_t) \, \partial_j L(\boldsymbol{\eta}_t).$$

Equivalently,

$$\boldsymbol{\eta}_{t+1} = \arg\min_{\boldsymbol{\eta}} \left\{ \langle \boldsymbol{\nabla}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}_t), \boldsymbol{\eta} - \boldsymbol{\eta}_t \rangle + \frac{1}{2\alpha_t} (\boldsymbol{\eta} - \boldsymbol{\eta}_t)^i g_{ij}(\boldsymbol{\eta}_t) (\boldsymbol{\eta} - \boldsymbol{\eta}_t)^j \right\}.$$

The metric is now intrinsic, but the step is still a local linear update in the primal coordinates. The natural-gradient direction is reparametrization invariant, while the finite-step update remains a local discretization.

For the exponential-family likelihood,

$$\boldsymbol{\eta}^i_{t+1} = \boldsymbol{\eta}^i_t - \alpha_t \, g^{ij}(\boldsymbol{\eta}_t) (\boldsymbol{\eta}^*_{t,j} - \bar{T}_j).$$

Here, we use the tangent plane of the statistical manifold as guide, but the linear approximation is not guaranteed to be accurate or respect parameter bounds.

1.3 Mirror gradient descent

Mirror gradient descent (MGD) replaces the Euclidean quadratic penalty by the Bregman divergence generated by A,

$$D_A(\boldsymbol{x}_i \| \boldsymbol{\eta}) = A(\boldsymbol{x}_i) - A(\boldsymbol{\eta}) - \langle \boldsymbol{\nabla} A(\boldsymbol{\eta}), \boldsymbol{x}_i - \boldsymbol{\eta} \rangle.$$

The update is

$$\boldsymbol{\eta}_{t+1} = \arg\min_{\boldsymbol{\eta}} \left\{ \langle \boldsymbol{\nabla}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}_t), \boldsymbol{\eta} - \boldsymbol{\eta}_t \rangle + \frac{1}{\alpha_t} D_A(\boldsymbol{\eta} \| \boldsymbol{\eta}_t) \right\}.$$

The first-order condition induced by the Bregman divergence is

$$\boldsymbol{\nabla} A(\boldsymbol{\eta}_{t+1}) = \boldsymbol{\nabla} A(\boldsymbol{\eta}_t) - \alpha_t \, \boldsymbol{\nabla}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}_t),$$

that is,

$$\boldsymbol{\eta}^*_{t+1} = \boldsymbol{\eta}^*_t - \alpha_t \, \boldsymbol{\nabla}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}_t).$$

Thus MGD induces a linear update in the dual coordinates $\boldsymbol{\eta}^*$. It uses the covector $\partial_i L$ with the correct index position, without inserting δ^{ij} or g^{ij} by hand. Of course, dual mean coordinates could be used without the information-geometric background. What we have obtained, however, is an answer to the question that ordinary GD cannot answer, namely, which coordinate system to choose. In the exponential family, dual coordinates are flat, hence geodesic projections are linear equations. Non-trivial problems arise via embedding curvature of curved exponential (sub-)families.

For the exponential-family likelihood one obtains

$$\boldsymbol{\eta}^*_{t+1} = \boldsymbol{\eta}^*_t - \alpha_t(\boldsymbol{\eta}^*_t - \bar{\boldsymbol{T}}) = (1 - \alpha_t)\boldsymbol{\eta}^*_t + \alpha_t \bar{\boldsymbol{T}}.$$

Hence MGD moves along the m-geodesic joining $\boldsymbol{\eta}^*_t$ and $\bar{\boldsymbol{T}}$, and the next primal iterate is

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\nabla} A^*(\boldsymbol{\eta}^*_{t+1}).$$

1.4 Boosting of MGD

We now formulate a boosting algorithm that approximates mirror gradient descent by updating the dual coordinate $\eta^*(x)$ directly. The construction separates three levels:

1. the population objective under the true data-generating distribution P,
2. the exact MGD flow in the dual coordinate η^*,
3. a weighted empirical tree algorithm based on unbiased pseudo-responses.

Throughout, expectations under the true distribution are written as $\mathbb{E}[\cdot]$, while weighted empirical averages are

$$\mathbb{E}_n^w[g] := \frac{1}{W}\sum_{i=1}^n w_i\, g_i, \qquad W := \sum_{i=1}^n w_i, \qquad w_i \geq 0.$$

Population optimum. Consider a conditional exponential family

$$p(y \mid x, \eta(x)) = h(y)\exp\big(\langle \eta(x), T(y)\rangle - A(\eta(x))\big),$$

with dual coordinate

$$\eta^*(x) = \nabla A(\eta(x)).$$

The population negative log-likelihood is

$$\mathcal{L}[\eta] = \mathbb{E}[A(\eta(X)) - \langle \eta(X), T(Y)\rangle - \log h(Y)]\,.$$

For an infinitely expressive model, the minimizer is pointwise and satisfies

$$\eta^*(x) = \mathbb{E}[T(Y) \mid X = x].$$

Thus the optimal dual coordinate is the conditional mean of the sufficient statistic under the true distribution.

Population MGD. At iteration t, bias the local objective by the Bregman divergence to the previous iterate:

$$\eta_{t+1}(x) = \arg\min_{\eta}\left\{A(\eta) - \langle \eta, \eta^*(x)\rangle + \frac{1}{\alpha}D_A(\eta\|\eta_t(x))\right\},$$

with constant learning rate $\alpha \in (0, 1]$. The first-order condition gives

$$\eta^*_{t+1}(x) = (1-\alpha)\eta^*_t(x) + \alpha\,\eta^*(x).$$

Equivalently,

$$\eta^*_{t+1}(x) - \eta^*_t(x) = \alpha(\eta^*(x) - \eta^*_t(x)).$$

So the exact mirror-descent direction in dual coordinates is

$$r_t(x) := \eta^*(x) - \eta^*_t(x).$$

Residual sufficient statistic as an unbiased target. Since

$$\eta^*(x) = \mathbb{E}[T(Y) \mid X = x],$$

the residual sufficient statistic

$$R_t := T(Y) - \hat{\eta}^*_t(X)$$

satisfies

$$\mathbb{E}[R_t \mid X = x] = \eta^*(x) - \hat{\eta}^*_t(x).$$

Thus R_t is an unbiased estimator of the exact dual MGD direction.

This is the key observation: one can approximate the population MGD step by learning the conditional mean of R_t and updating $\hat{\eta}^*_t$ additively.

Infinitely expressive regression class. If the regression class is rich enough to fit the conditional mean exactly,

$$f_{t+1}(x) = \mathbb{E}[R_t \mid X = x] = \eta^*(x) - \hat{\eta}_t^*(x),$$

then the additive update

$$\hat{\eta}_{t+1}^*(x) = \hat{\eta}_t^*(x) + \alpha\, f_{t+1}(x)$$

becomes

$$\hat{\eta}_{t+1}^*(x) = (1-\alpha)\hat{\eta}_t^*(x) + \alpha\, \eta^*(x),$$

which is exactly the population MGD recursion. Hence the construction is exact in the infinitely expressive limit.

Weighted empirical boosting algorithm. Let (x_i, y_i, w_i) be a weighted sample. Initialize

$$\hat{\eta}_0^*(x) = m_0,$$

with $m_0 \in \mathcal{N}^*$ any admissible baseline (often $m_0 = 0$ when $0 \in \mathcal{N}^*$). Define the weighted pseudo-response at round t by

$$R_i^{(t)} := T(y_i) - \hat{\eta}_t^*(x_i).$$

Then

$$\mathbb{E}[R_i^{(t)} \mid x_i] = \eta^*(x_i) - \hat{\eta}_t^*(x_i).$$

Fit a tree f_{t+1} to the pairs $(x_i, R_i^{(t)})$ with event weights w_i, using any regression criterion whose population minimizer is the conditional mean. The simplest choice is weighted least squares,

$$f_{t+1} = \arg\min_{f \in \mathcal{F}_{\text{tree}}} \sum_{i=1}^{n} w_i \, \|R_i^{(t)} - f(x_i)\|^2.$$

For a leaf J, the optimal constant prediction is the weighted mean residual

$$f_{t+1,J} = \frac{\sum_{i \in J} w_i \, R_i^{(t)}}{\sum_{i \in J} w_i}.$$

Update

$$\hat{\eta}_{t+1}^*(x) = \hat{\eta}_t^*(x) + \alpha\, f_{t+1}(x),$$

or, eventwise,

$$R_i^{(t+1)} = R_i^{(t)} - \alpha\, f_{t+1}(x_i).$$

After B rounds,

$$\hat{\eta}_B^*(x) = m_0 + \sum_{b=1}^{B} \alpha\, f_b(x), \qquad \hat{\eta}_B(x) = \nabla A^*(\hat{\eta}_B^*(x)).$$

Why this is not natural gradient descent in disguise. This construction does *not* reintroduce natural gradient descent through the back door. It starts from the exact mirror-descent recursion

$$\eta_{t+1}^* = (1-\alpha)\eta_t^* + \alpha\, \eta^*$$

in the distinguished coordinate system η^* selected by the Bregman geometry of the family. The residual target

$$T(Y) - \hat{\eta}_t^*(X)$$

is an unbiased estimator of the *dual* MGD direction itself. It is not, in general, the first term in a Taylor expansion of the exact nonlinear leafwise problem in the primal coordinate η, nor does it amount to Fisher-preconditioned gradient descent in η. Only in the infinitesimal-step regime do all first-order schemes coincide locally. Here the learning rate is applied directly in the dual coordinate η^*, and the update rule is globally valid throughout the exponential family because it is induced by the mirror map $\eta^* = \nabla A(\eta)$.

Split criterion from leaf predictions: empirical algorithm. The fitted leaf values immediately define the split criterion. Let J be a parent node and $J = L \cup R$ a candidate split. Define

$$W_C := \sum_{i \in C} w_i, \qquad \bar{R}_C^{(t)} := \frac{1}{W_C} \sum_{i \in C} w_i \, R_i^{(t)}, \qquad C \in \{J, L, R\}.$$

Then the optimal constant predictions are exactly

$$f_{t+1,C} = \bar{R}_C^{(t)}.$$

For weighted least squares, the reduction in within-node residual loss is

$$\mathrm{Gain}_t^{\mathrm{reg}}(J \to L, R) = W_L \|\bar{R}_L^{(t)}\|^2 + W_R \|\bar{R}_R^{(t)}\|^2 - W_J \|\bar{R}_J^{(t)}\|^2.$$

So the split can be chosen using the child leaf predictions alone. This is the direct vector-valued analogue of the usual CART variance reduction criterion.

Split criterion in the infinitely expressive limit. At population level, define

$$r_C^{(t)} := \mathbb{E}[R_t \mid X \in C] = \mathbb{E}[T(Y) - \eta_t^*(X) \mid X \in C], \qquad p_C := P_X(X \in C).$$

Then the best constant predictor on C is $r_C^{(t)}$, and the population split gain is

$$\mathrm{Gain}_{P,t}^{\mathrm{reg}}(J \to L, R) = p_L \|r_L^{(t)}\|^2 + p_R \|r_R^{(t)}\|^2 - p_J \|r_J^{(t)}\|^2.$$

Thus the empirical split rule above consistently estimates the corresponding population improvement in the dual MGD direction.

Entropy and the exact leafwise NLL criterion. There is a second, complementary viewpoint. If one fits a tree *exactly* by weighted negative log-likelihood on a fixed partition $\mathcal{P} = \{J\}$, then in each leaf

$$\eta_J^* = \frac{\sum_{i \in J} w_i \, T(y_i)}{\sum_{i \in J} w_i}.$$

Inserting this back gives, up to the carrier term,

$$\mathbb{E}_n^w[A(\eta(X)) - \langle \eta(X), T(Y) \rangle] = -\sum_{J \in \mathcal{P}} \frac{W_J}{W} \, A^*(\eta_J^*).$$

At population level the analogous expression is

$$-\sum_{J \in \mathcal{P}} P_X(J) \, A^*(\mathbb{E}[T(Y) \mid X \in J]).$$

So, for an exactly fitted tree, the optimized negative log-likelihood is a phase-space average of $-A^*(\eta^*)$. This is the generalized entropy story: $-A^*(\eta^*)$ is the non-carrier part of the entropy in the exponential family.

Accordingly, an exact leafwise split gain is

$$\mathrm{Gain}^{\mathrm{ent}}(J \to L, R) = \frac{W_L}{W} A^*(\eta_L^*) + \frac{W_R}{W} A^*(\eta_R^*) - \frac{W_J}{W} A^*(\eta_J^*)$$

empirically, and

$$\mathrm{Gain}_P^{\mathrm{ent}}(J \to L, R) = p_L A^*(\eta_L^*) + p_R A^*(\eta_R^*) - p_J A^*(\eta_J^*)$$

at population level, where

$$\eta_C^* = \mathbb{E}[T(Y) \mid X \in C].$$

This entropy gain is exact for a fully refit tree on a given partition and, in particular, for the first tree when no previous prediction is present. The residual-boosting algorithm above should be understood as a stagewise approximation to the global MGD flow, with $\mathrm{Gain}^{\mathrm{reg}}$ as the practical split rule and $\mathrm{Gain}^{\mathrm{ent}}$ as the exact refit criterion for a standalone tree or partition.

Monitoring during training. The natural global quantity to monitor for the boosted model itself is the weighted empirical negative log-likelihood

$$\widehat{\mathcal{L}}_n^w[\hat{\eta}_t] = \frac{1}{W}\sum_{i=1}^{n} w_i\Big[A(\hat{\eta}_t(x_i)) - \langle \hat{\eta}_t(x_i), T(y_i)\rangle - \log h(y_i)\Big].$$

This is the empirical proxy for the population risk $\mathcal{L}[\eta_t]$, and it directly measures predictive performance.

The entropy story provides an additional geometric diagnostic. For a fully refit partition $\mathcal{P}$, the optimized NLL is, up to the carrier term,

$$-\sum_{J\in\mathcal{P}} \frac{W_J}{W} A^*(\eta_J^*).$$

Thus one may also monitor the corresponding generalized empirical entropy

$$\widehat{\mathcal{H}}_{A,n}^w(\mathcal{P}) := -\sum_{J\in\mathcal{P}} \frac{W_J}{W} A^*(\eta_J^*).$$

For a boosted model this entropy quantity is no longer itself the training objective unless the leaves are fully refit, but it remains useful as an interpretable summary of how much conditional structure the partition captures.

Implementation summary for coding. Assume the following objects are available:

$$T(y), \qquad \nabla A^*(\cdot) \text{ to map } \eta^* \mapsto \eta, \qquad \{(x_i, y_i, w_i)\}_{i=1}^n, \qquad \alpha, \qquad \mathcal{F}_{\text{tree}}.$$

Then the MGD boosting loop is:

$$\hat{\eta}_0^*(x) = m_0,$$

for $t = 0, 1, \ldots,$

$$R_i^{(t)} = T(y_i) - \hat{\eta}_t^*(x_i),$$

fit a weighted regression tree to the targets $R_i^{(t)}$,

$$f_{t+1} = \arg\min_{f\in\mathcal{F}_{\text{tree}}} \sum_i w_i \, \|R_i^{(t)} - f(x_i)\|^2,$$

with leaf values

$$f_{t+1,J} = \frac{\sum_{i\in J} w_i \, R_i^{(t)}}{\sum_{i\in J} w_i},$$

update the prediction in the *dual* coordinate,

$$\hat{\eta}_{t+1}^*(x) = \hat{\eta}_t^*(x) + \alpha f_{t+1}(x),$$

optionally update residuals in place,

$$R_i^{(t+1)} = R_i^{(t)} - \alpha f_{t+1}(x_i),$$

and recover the natural coordinate only when needed,

$$\hat{\eta}_{t+1}(x) = \nabla A^*(\hat{\eta}_{t+1}^*(x)).$$

To choose splits in a node J, use

$$\text{Gain}_t^{\text{reg}}(J \to L, R) = W_L\|\bar{R}_L^{(t)}\|^2 + W_R\|\bar{R}_R^{(t)}\|^2 - W_J\|\bar{R}_J^{(t)}\|^2.$$

To monitor training globally, use

$$\widehat{\mathcal{L}}_n^w[\hat{\eta}_t] = \frac{1}{W}\sum_i w_i\Big[A(\hat{\eta}_t(x_i)) - \langle \hat{\eta}_t(x_i), T(y_i)\rangle - \log h(y_i)\Big].$$

1.5 Boosting of NGD

Natural-gradient descent suggests a second stagewise boosting construction, now based on updating the natural coordinate $\eta(x)$ itself. In contrast to MGD, this construction is local: the Fisher metric is evaluated at the current iterate and used to precondition the sufficient-statistic residual.

We again consider the conditional exponential family

$$p(y \mid x, \eta(x)) = h(y) \exp(\langle \eta(x), T(y) \rangle - A(\eta(x))),$$

with dual coordinate

$$\eta^*(x) = \nabla A(\eta(x)).$$

The population negative log-likelihood is

$$\mathcal{L}_P[\eta] = \mathbb{E}_P[A(\eta(X)) - \langle \eta(X), T(Y) \rangle - \log h(Y)] .$$

Its pointwise Euclidean gradient in the natural parameter is

$$\nabla_\eta \mathcal{L}_P(x) = \eta^*(x) - \mathbb{E}_P[T(Y) \mid X = x].$$

The Fisher metric of the family is

$$g(\eta) = \nabla^2 A(\eta).$$

Hence the negative natural-gradient direction is

$$v_t(x) = -\, g(\eta_t(x))^{-1} \nabla_\eta \mathcal{L}_P(x) = g(\eta_t(x))^{-1} \Big(\mathbb{E}_P[T(Y) \mid X = x] - \eta_t^*(x) \Big).$$

The population NGD Euler step is therefore

$$\eta_{t+1}(x) = \eta_t(x) + \alpha\, v_t(x).$$

Residual sufficient statistic as an unbiased natural-gradient target. Define

$$U_t := g(\hat{\eta}_t(X))^{-1} \Big(T(Y) - \hat{\eta}_t^*(X) \Big).$$

Then

$$\mathbb{E}_P[U_t \mid X = x] = g(\hat{\eta}_t(x))^{-1} \Big(\mathbb{E}_P[T(Y) \mid X = x] - \hat{\eta}_t^*(x) \Big) = v_t(x).$$

Thus U_t is an unbiased estimator of the population NGD direction.

If the regression class is infinitely expressive and can fit the conditional mean exactly, then

$$f_{t+1}(x) = \mathbb{E}_P[U_t \mid X = x] = v_t(x)$$

and the stagewise update

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha\, f_{t+1}(x)$$

reproduces the exact population NGD step.

Weighted empirical algorithm. Let (x_i, y_i, w_i) be a weighted sample and define

$$W := \sum_{i=1}^{n} w_i, \qquad \hat{\eta}_t^*(x_i) = \nabla A(\hat{\eta}_t(x_i)), \qquad g_i^{(t)} := g(\hat{\eta}_t(x_i)).$$

The empirical NGD pseudo-response is

$$U_i^{(t)} = (g_i^{(t)})^{-1}\Big(T(y_i) - \hat{\eta}_t^*(x_i)\Big).$$

Fit a regression tree f_{t+1} to the pairs $(x_i, U_i^{(t)})$ with event weights w_i. The simplest choice is weighted least squares:

$$f_{t+1} = \arg\min_{f \in \mathcal{F}_{\text{tree}}} \sum_{i=1}^{n} w_i \, \|U_i^{(t)} - f(x_i)\|^2.$$

For a leaf J, the optimal constant leaf value is

$$f_{t+1,J} = \frac{\sum_{i \in J} w_i \, U_i^{(t)}}{\sum_{i \in J} w_i}.$$

The update is then

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha \, f_{t+1}(x).$$

The corresponding dual prediction must then be recomputed through the Legendre map:

$$\hat{\eta}_{t+1}^*(x) = \nabla A(\hat{\eta}_{t+1}(x)).$$

In contrast to MGD, the residual shift is therefore not linear in the fitted tree. The exact updated sufficient-statistic residual is

$$R_{t+1} = T(Y) - \hat{\eta}_{t+1}^*(X) = T(Y) - \nabla A\big(\hat{\eta}_t(X) + \alpha f_{t+1}(X)\big),$$

and the next unbiased NGD target is obtained by recomputing

$$U_{t+1} = g(\hat{\eta}_{t+1}(X))^{-1} R_{t+1}.$$

Leafwise split criterion. Define

$$W_C := \sum_{i \in C} w_i, \qquad \bar{U}_C^{(t)} := \frac{\sum_{i \in C} w_i \, U_i^{(t)}}{W_C}, \qquad C \in \{J, L, R\}.$$

Then the usual weighted regression gain is

$$\text{Gain}_t^{\text{NGD}}(J \to L, R) = W_L \|\bar{U}_L^{(t)}\|^2 + W_R \|\bar{U}_R^{(t)}\|^2 - W_J \|\bar{U}_J^{(t)}\|^2.$$

Thus NGD boosting uses the same tree machinery as ordinary weighted regression, but on a pseudo-response that has been preconditioned by the inverse Fisher metric.

Local versus global geometry. MGD becomes linear in the global dual coordinate η^*, while NGD remains a local tangent-plane approximation in the natural coordinate η. Indeed,

$$\hat{\eta}_{t+1}^* = \nabla A(\hat{\eta}_t + \alpha v_t) = \hat{\eta}_t^* + \alpha \, g(\hat{\eta}_t) v_t + O(\alpha^2) = \hat{\eta}_t^* + \alpha(\eta^* - \hat{\eta}_t^*) + O(\alpha^2),$$

so NGD and MGD agree to first order in α, but differ at finite step size. In this sense, MGD is globally adapted to the exponential-family geometry, whereas NGD is its local quadratic approximation.

Monitoring during training. As for MGD, the natural global quantity to monitor is the weighted empirical negative log-likelihood

$$\widehat{\mathcal{L}}_n^w[\hat{\eta}_t] = \frac{1}{W}\sum_{i=1}^{n} w_i\Big[A(\hat{\eta}_t(x_i)) - \langle \hat{\eta}_t(x_i), T(y_i)\rangle - \log h(y_i)\Big].$$

This directly measures predictive performance. One may again compare it to the exact partition-level entropy criterion

$$\widehat{\mathcal{H}}_{A,n}^w(\mathcal{P}) = -\sum_{J\in\mathcal{P}} \frac{W_J}{W} A^*(\eta_J^*), \qquad \eta_J^* = \frac{\sum_{i\in J} w_i\, T(y_i)}{\sum_{i\in J} w_i},$$

but for NGD boosting this entropy is an interpretive diagnostic, not the stagewise training objective.

Implementation summary for coding. Assume the following objects are available:

$$T(y), \qquad \nabla A(\cdot) \text{ to map } \eta \mapsto \eta^*, \qquad \nabla^2 A(\cdot) \text{ to compute } g(\eta), \qquad \{(x_i, y_i, w_i)\}_{i=1}^n, \qquad \alpha, \qquad \mathcal{F}_{\text{tree}}.$$

Then the NGD boosting loop is: initialize a natural-coordinate model $\hat{\eta}_0(x)$, compute

$$\hat{\eta}_0^*(x_i) = \nabla A(\hat{\eta}_0(x_i)), \qquad g_i^{(0)} = \nabla^2 A(\hat{\eta}_0(x_i)),$$

and for $t = 0, 1, \ldots,$

$$U_i^{(t)} = (g_i^{(t)})^{-1}(T(y_i) - \hat{\eta}_t^*(x_i)),$$

fit a weighted regression tree to the targets $U_i^{(t)}$,

$$f_{t+1} = \arg\min_{f\in\mathcal{F}_{\text{tree}}} \sum_i w_i \,\|U_i^{(t)} - f(x_i)\|^2,$$

with leaf values

$$f_{t+1,J} = \frac{\sum_{i\in J} w_i\, U_i^{(t)}}{\sum_{i\in J} w_i},$$

update the prediction in the *natural* coordinate,

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha f_{t+1}(x),$$

then recompute

$$\hat{\eta}_{t+1}^*(x_i) = \nabla A(\hat{\eta}_{t+1}(x_i)), \qquad g_i^{(t+1)} = \nabla^2 A(\hat{\eta}_{t+1}(x_i)).$$

To choose splits in a node J, use

$$\text{Gain}_t^{\text{NGD}}(J \to L, R) = W_L\|\bar{U}_L^{(t)}\|^2 + W_R\|\bar{U}_R^{(t)}\|^2 - W_J\|\bar{U}_J^{(t)}\|^2.$$

To monitor training globally, use

$$\widehat{\mathcal{L}}_n^w[\hat{\eta}_t] = \frac{1}{W}\sum_{i} w_i\Big[A(\hat{\eta}_t(x_i)) - \langle \hat{\eta}_t(x_i), T(y_i)\rangle - \log h(y_i)\Big].$$

Algorithmic simplification: recovering MGD from an NGD implementation. If an NGD implementation already exists, then the tree-fitting machinery can be reused almost verbatim for MGD. What changes is the coordinate bookkeeping. In NGD, one forms

$$U_i^{(t)} = (g_i^{(t)})^{-1}(T(y_i) - \hat{\eta}_t^*(x_i)),$$

fits a tree in the natural coordinate, updates $\hat{\eta}_t$, and recomputes $\hat{\eta}_t^* = \nabla A(\hat{\eta}_t)$. To obtain MGD instead, one drops the Fisher-preconditioning factor g_i^{-1}, works directly with

$$R_i^{(t)} = T(y_i) - \hat{\eta}_t^*(x_i),$$

fits the tree in the *dual* coordinate, updates

$$\hat{\eta}_{t+1}^* = \hat{\eta}_t^* + \alpha f_{t+1},$$

and maps back only when needed through

$$\hat{\eta}_{t+1} = \nabla A^*(\hat{\eta}_{t+1}^*).$$

Equivalently: starting from an NGD code path, MGD is obtained by removing Fisher preconditioning, changing the fitted coordinate from η to η^*, and replacing the nonlinear residual recomputation by the exact additive dual-residual update

$$R_i^{(t+1)} = R_i^{(t)} - \alpha f_{t+1}(x_i).$$

So MGD is not literally the same algorithm as NGD, but algorithmically it is the simpler special case once the exponential-family maps are available.

1.6 Primal–dual pairs and boosted realizations

The discussion above shows that, at the level of pointwise optimization in an unrestricted function space, natural-gradient descent and mirror-gradient descent come in two primal–dual pairs. Boosting changes this picture because the model class is no longer closed under the nonlinear Legendre map $\eta^* = \nabla A(\eta)$. This yields four practically distinct boosted algorithms: two *matched* constructions, where the additive tree expansion is written in the coordinate in which the corresponding update law is simple, and two *crossed* constructions, where the update law is represented in the other coordinate.

1.6.1 Two pointwise update laws and their primal–dual pairings

Let $F(\eta)$ be a smooth objective in the natural coordinate, and let

$$\widetilde{F}(\eta^*) := F(\nabla A^*(\eta^*))$$

be its reparametrization in the dual coordinate. Since

$$g(\eta) = \nabla^2 A(\eta), \qquad g^*(\eta^*) = \nabla^2 A^*(\eta^*) = g(\eta)^{-1},$$

the chain rule gives

$$\nabla_{\eta^*}\widetilde{F}(\eta^*) = g(\eta)^{-1}\nabla_\eta F(\eta).$$

There are then two distinct pointwise update laws.

The M-law. Mirror descent in the primal coordinate η with mirror potential A gives

$$\eta_+^* = \eta^* - \alpha\, \nabla_\eta F(\eta).$$

Equivalently, natural gradient descent in the dual coordinate η^* gives

$$\eta_+^* = \eta^* - \alpha\, (g^*(\eta^*))^{-1} \nabla_{\eta^*} \widetilde{F}(\eta^*) = \eta^* - \alpha\, \nabla_\eta F(\eta).$$

Thus

$$\boxed{\text{primal MGD} = \text{dual NGD.}}$$

This is the update law that is linear in the dual coordinate η^*.

The N-law. Natural gradient descent in the primal coordinate η gives

$$\eta_+ = \eta - \alpha\, g(\eta)^{-1} \nabla_\eta F(\eta).$$

Equivalently, mirror descent in the dual coordinate η^* with mirror potential A^* gives

$$\eta_+ = \eta - \alpha\, \nabla_{\eta^*} \widetilde{F}(\eta^*) = \eta - \alpha\, g(\eta)^{-1} \nabla_\eta F(\eta).$$

Thus

$$\boxed{\text{primal NGD} = \text{dual MGD.}}$$

This is the update law that is linear in the natural coordinate η.

Exponential-family negative log-likelihood. For

$$F(\eta) = A(\eta) - \langle \eta, \bar{T} \rangle, \qquad \nabla_\eta F(\eta) = \eta^* - \bar{T},$$

the M-law becomes

$$\eta_+^* = (1 - \alpha)\eta^* + \alpha \bar{T},$$

while the N-law becomes

$$\eta_+ = \eta - \alpha\, g(\eta)^{-1} (\eta^* - \bar{T}).$$

These are the two pointwise update laws that underlie all four boosted algorithms below.

1.6.2 Matched and crossed boosted algorithms

The earlier subsections introduced the two *matched* boosted realizations: MGD boosting for the M-law, written as an additive expansion in the dual coordinate η^*, and NGD boosting for the N-law, written as an additive expansion in the natural coordinate η.

The reason these are the natural constructions is that the corresponding update laws are simple in those coordinates:

$$M\text{-law: linear in } \eta^*, \qquad N\text{-law: linear in } \eta.$$

Boosting breaks the primal–dual equivalence because the model class is restricted. Indeed, an additive tree model in η,

$$\hat{\eta}_B(x) = \hat{\eta}_0(x) + \sum_{b=1}^{B} \alpha\, g_b(x),$$

is in general not mapped by ∇A to an additive tree model in η^*, and conversely an additive tree model in η^*,

$$\hat{\eta}_B^*(x) = \hat{\eta}_0^*(x) + \sum_{b=1}^{B} \alpha\, f_b(x),$$

is in general not mapped by ∇A^* to an additive tree model in η. Hence each pointwise update law admits two distinct boosted realizations.

Matched realization of the M-law: MGD boosting in η^*. This is the construction developed above:

$$\hat{\eta}^*_{t+1}(x) = \hat{\eta}^*_t(x) + \alpha f_{t+1}(x),$$

with residual sufficient-statistic target

$$R_t = T(Y) - \hat{\eta}^*_t(X), \qquad \mathbb{E}[R_t \mid X = x] = \eta^*(x) - \hat{\eta}^*_t(x).$$

This realization is exact in the infinitely expressive limit and admits the exact in-place residual recursion

$$R_{t+1} = R_t - \alpha f_{t+1}(X).$$

Matched realization of the N-law: NGD boosting in η. This is the second construction developed above:

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha g_{t+1}(x),$$

with Fisher-preconditioned pseudo-response

$$U_t = g(\hat{\eta}_t(X))^{-1}\big(T(Y) - \hat{\eta}^*_t(X)\big), \qquad \mathbb{E}[U_t \mid X = x] = g(\hat{\eta}_t(x))^{-1}\big(\eta^*(x) - \hat{\eta}^*_t(x)\big).$$

This realization is exact in the infinitely expressive limit for the N-law, but its residual transport is nonlinear because after the primal update one must recompute

$$\hat{\eta}^*_{t+1}(x) = \nabla A(\hat{\eta}_{t+1}(x)).$$

Crossed realization of the M-law: additive trees in η. Here the ideal pointwise target is still the M-law

$$\tilde{\eta}^*_{t+1}(x) = (1-\alpha)\hat{\eta}^*_t(x) + \alpha\, \eta^*(x),$$

but the boosted model is represented additively in the natural coordinate:

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha g_{t+1}(x).$$

The corresponding ideal primal increment is therefore

$$q_t^{(M,\eta)}(x) := \frac{1}{\alpha}\left[\nabla A^*\Big((1-\alpha)\hat{\eta}^*_t(x) + \alpha\, \eta^*(x)\Big) - \hat{\eta}_t(x)\right].$$

If one could fit $q_t^{(M,\eta)}(x)$ exactly, this would realize the M-law in a primal additive model. In general, however, there is no simple unbiased sample-level pseudo-response for $q_t^{(M,\eta)}$, precisely because the map ∇A^* is nonlinear. This is why the matched dual-coordinate realization is algorithmically much cleaner.

Crossed realization of the N-law: additive trees in η^*. Here the ideal pointwise target is the N-law

$$\tilde{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha\, g(\hat{\eta}_t(x))^{-1}\big(\eta^*(x) - \hat{\eta}^*_t(x)\big),$$

but the boosted model is represented additively in the dual coordinate:

$$\hat{\eta}^*_{t+1}(x) = \hat{\eta}^*_t(x) + \alpha f_{t+1}(x).$$

The corresponding ideal dual increment is

$$q_t^{(N,\eta^*)}(x) := \frac{1}{\alpha}\left[\nabla A\Big(\hat{\eta}_t(x) + \alpha\, g(\hat{\eta}_t(x))^{-1}\big(\eta^*(x) - \hat{\eta}^*_t(x)\big)\Big) - \hat{\eta}^*_t(x)\right].$$

Again, if one could fit $q_t^{(N,\eta^*)}(x)$ exactly, this would realize the N-law in a dual additive model. In general there is no simple unbiased pseudo-response because of the nonlinear map ∇A. This is the dual counterpart of the previous crossed construction.

Consequences. Thus the unrestricted pointwise duality collapses to four distinct boosted algorithms:

1. M-law in the dual coordinate η^* (matched; earlier MGD boosting),
2. N-law in the natural coordinate η (matched; earlier NGD boosting),
3. M-law in the natural coordinate η (crossed),
4. N-law in the dual coordinate η^* (crossed).

The matched versions are the natural ones because they preserve the simple linear structure of the corresponding pointwise update laws. The crossed versions are meaningful, but they lose the simple unbiased pseudo-responses and exact residual recursions that made the matched algorithms attractive.

1.6.3 Poisson example

For the Poisson family,

$$A(\eta) = e^{\eta}, \qquad \eta^* = \mu = e^{\eta}, \qquad \eta = \log \mu, \qquad g(\eta) = \mu.$$

The two pointwise laws are

$$M\text{-law:} \quad \mu_+ = (1-\alpha)\mu + \alpha \bar{T},$$

and

$$N\text{-law:} \quad \eta_+ = \eta - \alpha \, \frac{\mu - \bar{T}}{\mu}.$$

In Poisson language $\bar{T}$ is just the target mean.

The four boosted realizations are then all distinct.

(i) M-law in μ: matched MGD boosting. The model is additive in μ:

$$\hat{\mu}_{t+1}(x) = \hat{\mu}_t(x) + \alpha f_{t+1}(x),$$

with unbiased target

$$R_t = Y - \hat{\mu}_t(X), \qquad \mathbb{E}[R_t \mid X = x] = \mu(x) - \hat{\mu}_t(x).$$

This is the clean affine residual recursion.

(ii) N-law in $\eta = \log \mu$: matched NGD boosting. The model is additive in η:

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha g_{t+1}(x),$$

with unbiased target

$$U_t = \frac{Y - \hat{\mu}_t(X)}{\hat{\mu}_t(X)} = \frac{Y}{\hat{\mu}_t(X)} - 1,$$

since $g(\eta) = \mu$. This is the natural-gradient pseudo-response.

(iii) M-law in $\eta = \log \mu$: crossed realization. The ideal next mean is

$$\tilde{\mu}_{t+1}(x) = (1-\alpha)\hat{\mu}_t(x) + \alpha \, \mu(x),$$

but the tree is additive in η, so the ideal primal increment is

$$q_t^{(M,\eta)}(x) = \frac{1}{\alpha} \left[\log \left((1-\alpha)\hat{\mu}_t(x) + \alpha \, \mu(x) \right) - \log \hat{\mu}_t(x) \right].$$

This is different from the matched NGD target $(\mu - \hat{\mu}_t)/\hat{\mu}_t$.

(iv) N-law in μ: crossed realization. The ideal primal update is

$$\tilde{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha \frac{\mu(x) - \hat{\mu}_t(x)}{\hat{\mu}_t(x)},$$

so the corresponding ideal dual increment is

$$q_t^{(N,\mu)}(x) = \frac{1}{\alpha}\left[\hat{\mu}_t(x) \exp\left(\alpha \frac{\mu(x) - \hat{\mu}_t(x)}{\hat{\mu}_t(x)}\right) - \hat{\mu}_t(x)\right].$$

This is different from the matched MGD residual $\mu - \hat{\mu}_t$.

Hence even in the simplest non-Gaussian case, Poisson, the four boosted realizations are all distinct. Only in the Gaussian family, where the map $\eta \leftrightarrow \eta^*$ is affine and the Fisher metric is constant, do the distinctions collapse.

1.7 NEF–QVF specializations of the four boosted update laws

We now specialize the four boosted update laws to the six one-dimensional Morris families. In one dimension we write

$$\mu := \eta^* = A'(\eta), \qquad \eta = \eta(\mu) = A^{*\prime}(\mu), \qquad V(\mu) = A''(\eta(\mu)).$$

At a fixed boosting round t, let

$$\hat{\mu} = \hat{\eta}_t^*(x), \qquad \hat{\eta} = \eta(\hat{\mu}), \qquad m = \eta^*(x) = \mathbb{E}[T(Y) \mid X = x].$$

The learning rate is $\alpha \in (0, 1]$.

Generic implementation templates. There are four boosted realizations.

(1) M-law in μ (matched MGD). The pointwise target law is

$$\mu_+ = (1 - \alpha)\hat{\mu} + \alpha m,$$

so the increment to be learned is

$$q_t^{(M,\mu)}(\hat{\mu}, m) = m - \hat{\mu}.$$

This admits the exact unbiased sample-level target

$$R_i^{(t)} = T(y_i) - \hat{\mu}_i.$$

A tree is fitted to $R_i^{(t)}$, the model is updated additively in μ,

$$\hat{\mu}_{t+1}(x) = \hat{\mu}_t(x) + \alpha f_{t+1}(x),$$

and then mapped back through $\hat{\eta}_{t+1}(x) = \eta(\hat{\mu}_{t+1}(x))$.

(2) N-law in η (matched NGD). The pointwise target law is

$$\eta_+ = \hat{\eta} + \alpha \frac{m - \hat{\mu}}{V(\hat{\mu})},$$

so the increment to be learned is

$$q_t^{(N,\eta)}(\hat{\mu}, m) = \frac{m - \hat{\mu}}{V(\hat{\mu})}.$$

This admits the exact unbiased sample-level target

$$U_i^{(t)} = \frac{T(y_i) - \hat{\mu}_i}{V(\hat{\mu}_i)}.$$

A tree is fitted to $U_i^{(t)}$, the model is updated additively in η,

$$\hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha g_{t+1}(x),$$

and then mapped forward through $\hat{\mu}_{t+1}(x) = \mu(\hat{\eta}_{t+1}(x))$.

(3) M-law in η (crossed realization). The pointwise target law is still

$$\mu_+ = (1-\alpha)\hat{\mu} + \alpha m,$$

but the boosted model is additive in η. Hence the ideal primal increment is

$$q_t^{(M,\eta)}(\hat{\mu}, m) = \frac{1}{\alpha}\Big[\eta((1-\alpha)\hat{\mu} + \alpha m) - \eta(\hat{\mu})\Big].$$

This is an oracle population target. In general there is no exact unbiased eventwise pseudo-response obtained by replacing m with $T(y)$, because $\eta(\mu)$ is nonlinear.

(4) N-law in μ (crossed realization). The pointwise target law is still

$$\eta_+ = \hat{\eta} + \alpha\,\frac{m - \hat{\mu}}{V(\hat{\mu})},$$

but the boosted model is additive in μ. Hence the ideal dual increment is

$$q_t^{(N,\mu)}(\hat{\mu}, m) = \frac{1}{\alpha}\left[\mu\left(\eta(\hat{\mu}) + \alpha\,\frac{m - \hat{\mu}}{V(\hat{\mu})}\right) - \hat{\mu}\right].$$

Again, this is an oracle population target and there is no exact unbiased eventwise pseudo-response in general.

Core family data. All six families are specified by the cumulant $A(\eta)$, the gradient map $\mu(\eta)$, its inverse $\eta(\mu)$, the variance function $V(\mu)$, and the dual potential $A^*(\mu)$ (up to an additive constant). These are collected in Table 1.

M-law in μ: matched MGD boosting. The generic implementation is the same for all six families:

$$R_i^{(t)} = T(y_i) - \hat{\mu}_i, \qquad f_{t+1,J} = \frac{\sum_{i\in J} w_i\, R_i^{(t)}}{\sum_{i\in J} w_i}, \qquad \hat{\mu}_{t+1}(x) = \hat{\mu}_t(x) + \alpha f_{t+1}(x).$$

The only family-specific ingredients are the admissible domain of μ, the map-back $\eta(\mu)$, and, if desired, the dual potential $A^*(\mu)$ for entropy monitoring. These are listed in Table 2.

N-law in η: matched NGD boosting. The generic implementation is again uniform:

$$U_i^{(t)} = \frac{T(y_i) - \hat{\mu}_i}{V(\hat{\mu}_i)}, \qquad g_{t+1,J} = \frac{\sum_{i\in J} w_i\, U_i^{(t)}}{\sum_{i\in J} w_i}, \qquad \hat{\eta}_{t+1}(x) = \hat{\eta}_t(x) + \alpha g_{t+1}(x),$$

followed by $\hat{\mu}_{t+1}(x) = \mu(\hat{\eta}_{t+1}(x))$. The family-specific ingredients are the forward map $\mu(\eta)$, the variance function $V(\mu)$, and the resulting eventwise target $U_i^{(t)}$. These are listed in Table 3.

Family	$A(\eta)$	$\mu(\eta)$	$\eta(\mu)$	$V(\mu)$	$A^*(\mu)$
Normal	$\frac{\sigma_0^2\eta^2}{2}$	$\sigma_0^2\eta$	$\frac{\mu}{\sigma_0^2}$	σ_0^2	$\frac{\mu^2}{2\sigma_0^2}$
Poisson	e^η	e^η	$\log\mu$	μ	$\mu\log\mu-\mu$
Gamma	$-r\log(-\eta)$	$\frac{r}{-\eta}$	$-\frac{r}{\mu}$	$\frac{\mu^2}{r}$	$-r\log\mu$
Binomial	$N\log(1+e^\eta)$	$N\frac{e^\eta}{1+e^\eta}$	$\log\frac{\mu}{N-\mu}$	$\mu\left(1-\frac{\mu}{N}\right)$	$\mu\log\frac{\mu}{N}+(N-\mu)\log\frac{N-\mu}{N}$
Neg. binomial	$-r\log(1-e^\eta)$	$\frac{re^\eta}{1-e^\eta}$	$\log\frac{\mu}{r+\mu}$	$\mu\left(1+\frac{\mu}{r}\right)$	$\mu\log\mu-(r+\mu)\log(r+\mu)$
GHS	$-2r\log(2\cos(\eta/2))$	$r\tan(\eta/2)$	$2\arctan(\mu/r)$	$\frac{r^2+\mu^2}{2r}$	$2\mu\arctan(\mu/r)-r\log(r^2+\mu^2)$

Table 1: Core one-dimensional NEF–QVF family data, up to irrelevant additive constants in $A(\eta)$ and $A^*(\mu)$.

Family	admissible μ	$\eta(\mu)$	$A^*(\mu)$
Normal	$\mu\in\mathbb{R}$	$\frac{\mu}{\sigma_0^2}$	$\frac{\mu^2}{2\sigma_0^2}$
Poisson	$\mu>0$	$\log\mu$	$\mu\log\mu-\mu$
Gamma	$\mu>0$	$-\frac{r}{\mu}$	$-r\log\mu$
Binomial	$0<\mu<N$	$\log\frac{\mu}{N-\mu}$	$\mu\log\frac{\mu}{N}+(N-\mu)\log\frac{N-\mu}{N}$
Negative binomial	$\mu>0$	$\log\frac{\mu}{r+\mu}$	$\mu\log\mu-(r+\mu)\log(r+\mu)$
GHS	$\mu\in\mathbb{R}$	$2\arctan(\mu/r)$	$2\mu\arctan(\mu/r)-r\log(r^2+\mu^2)$

Table 2: Family-specific plug-ins for matched MGD boosting in the dual coordinate $\mu=\eta^*$. The generic eventwise target is $R_i^{(t)}=T(y_i)-\hat{\mu}_i$.

M-law in η: crossed realization. The oracle pointwise target is

$$q_t^{(M,\eta)}(\hat{\mu},m)=\frac{1}{\alpha}\Big[\eta\big((1-\alpha)\hat{\mu}+\alpha m\big)-\eta(\hat{\mu})\Big].$$

This is the exact increment required to realize the affine M-law in a primal additive tree model. The family-specific formulas are listed in Table 4. In general there is no exact unbiased eventwise pseudo-response obtained by replacing m with $T(y)$.

N-law in μ: crossed realization. The oracle pointwise target is

$$q_t^{(N,\mu)}(\hat{\mu},m)=\frac{1}{\alpha}\left[\mu\left(\eta(\hat{\mu})+\alpha\frac{m-\hat{\mu}}{V(\hat{\mu})}\right)-\hat{\mu}\right].$$

This is the exact increment required to realize the N-law in a dual additive tree model. The family-specific formulas are listed in Table 5. Again, there is no exact unbiased eventwise pseudo-response in general.

Practical interpretation. The two matched realizations are the algorithmically clean ones:

- M-law in μ: additive trees in the dual coordinate with exact residual target $R_i=T(y_i)-\hat{\mu}_i$,

Family	$\mu(\eta)$	$V(\mu)$	$U_i^{(t)}$
Normal	$\sigma_0^2\eta$	σ_0^2	$\dfrac{T(y_i)-\hat{\mu}_i}{\sigma_0^2}$
Poisson	e^{η}	μ	$\dfrac{T(y_i)-\hat{\mu}_i}{\hat{\mu}_i}$
Gamma	$\dfrac{r}{-\eta}$	$\dfrac{\mu^2}{r}$	$\dfrac{r\,(T(y_i)-\hat{\mu}_i)}{\hat{\mu}_i^2}$
Binomial	$N\dfrac{e^{\eta}}{1+e^{\eta}}$	$\mu\left(1-\dfrac{\mu}{N}\right)$	$\dfrac{N\,(T(y_i)-\hat{\mu}_i)}{\hat{\mu}_i(N-\hat{\mu}_i)}$
Negative binomial	$\dfrac{re^{\eta}}{1-e^{\eta}}$	$\mu\left(1+\dfrac{\mu}{r}\right)$	$\dfrac{r\,(T(y_i)-\hat{\mu}_i)}{\hat{\mu}_i(r+\hat{\mu}_i)}$
GHS	$r\tan(\eta/2)$	$\dfrac{r^2+\mu^2}{2r}$	$\dfrac{2r\,(T(y_i)-\hat{\mu}_i)}{r^2+\hat{\mu}_i^2}$

Table 3: Family-specific plug-ins for matched NGD boosting in the natural coordinate η.

Family	$q_t^{(M,\eta)}(\hat{\mu}, m)$
Normal	$\dfrac{m-\hat{\mu}}{\sigma_0^2}$
Poisson	$\dfrac{1}{\alpha}\log\dfrac{(1-\alpha)\hat{\mu}+\alpha m}{\hat{\mu}}$
Gamma	$\dfrac{r(m-\hat{\mu})}{\hat{\mu}((1-\alpha)\hat{\mu}+\alpha m)}$
Binomial	$\dfrac{1}{\alpha}\log\dfrac{((1-\alpha)\hat{\mu}+\alpha m)(N-\hat{\mu})}{\hat{\mu}(N-(1-\alpha)\hat{\mu}-\alpha m)}$
Negative binomial	$\dfrac{1}{\alpha}\log\dfrac{((1-\alpha)\hat{\mu}+\alpha m)(r+\hat{\mu})}{\hat{\mu}(r+(1-\alpha)\hat{\mu}+\alpha m)}$
GHS	$\dfrac{2}{\alpha}\left[\arctan\left(\dfrac{(1-\alpha)\hat{\mu}+\alpha m}{r}\right)-\arctan\left(\dfrac{\hat{\mu}}{r}\right)\right]$

Table 4: Oracle pointwise targets for the crossed realization of the M-law in the natural coordinate η.

- N-law in η: additive trees in the natural coordinate with exact Fisher-preconditioned target $U_i = (T(y_i) - \hat{\mu}_i)/V(\hat{\mu}_i)$.

The two crossed realizations are mathematically well defined and are listed above for completeness, but they no longer admit the same simple unbiased eventwise pseudo-responses because the nonlinear Legendre map breaks the additive tree structure.

1.8 Curved exponential families

A curved exponential family is a lower-dimensional submanifold of the ambient family, parameterized by

$$\psi \in \Psi \subset \mathbb{R}^m, \qquad m < d,$$

through an embedding

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\psi) \in \mathcal{N}.$$

Its dual embedding is

$$\boldsymbol{\eta}^*(\psi) = \boldsymbol{\nabla} A(\boldsymbol{\eta}(\psi)) \in \mathcal{N}^*.$$

Family	$q_t^{(N,\mu)}(\hat{\mu}, m)$
Normal	$m - \hat{\mu}$
Poisson	$\frac{\hat{\mu}}{\alpha}\left[\exp\left(\alpha\frac{m-\hat{\mu}}{\hat{\mu}}\right) - 1\right]$
Gamma	$\frac{\hat{\mu}(m-\hat{\mu})}{(1+\alpha)\hat{\mu} - \alpha m}$
Binomial	$\frac{1}{\alpha}\left[\frac{N\rho}{1+\rho} - \hat{\mu}\right], \quad \rho = \frac{\hat{\mu}}{N-\hat{\mu}}\exp\left(\alpha\frac{N(m-\hat{\mu})}{\hat{\mu}(N-\hat{\mu})}\right)$
Negative binomial	$\frac{1}{\alpha}\left[\frac{r\rho}{1-\rho} - \hat{\mu}\right], \quad \rho = \frac{\hat{\mu}}{r+\hat{\mu}}\exp\left(\alpha\frac{r(m-\hat{\mu})}{\hat{\mu}(r+\hat{\mu})}\right)$
GHS	$\frac{1}{\alpha}\left[r\tan\left(\arctan(\hat{\mu}/r) + \alpha\,\frac{r(m-\hat{\mu})}{r^2+\hat{\mu}^2}\right) - \hat{\mu}\right]$

Table 5: Oracle pointwise targets for the crossed realization of the N-law in the dual coordinate μ.

The curved negative log-likelihood is

$$L(\psi) = A(\boldsymbol{\eta}(\psi)) - \langle \boldsymbol{\eta}(\psi), \bar{\boldsymbol{T}} \rangle.$$

At iteration t, let

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}(\psi_t), \qquad \boldsymbol{\eta}_t^* = \boldsymbol{\eta}^*(\psi_t),$$

and define the ambient mirror target

$$\tilde{\boldsymbol{\eta}}_t^* = (1-\alpha_t)\boldsymbol{\eta}_t^* + \alpha_t \bar{\boldsymbol{T}}, \qquad \tilde{\boldsymbol{\eta}}_t = \boldsymbol{\nabla} A^*(\tilde{\boldsymbol{\eta}}_t^*).$$

Thus $\tilde{\boldsymbol{\eta}}_t^*$ lies on the m-geodesic segment joining $\boldsymbol{\eta}_t^*$ and $\bar{\boldsymbol{T}}$. Since $\tilde{\boldsymbol{\eta}}_t$ need not lie on the curved model, the curved mirror step is the constrained Bregman projection

$$\boxed{\psi_{t+1} = \arg\min_{\psi\in\Psi} D_A(\boldsymbol{\eta}(\psi)\|\tilde{\boldsymbol{\eta}}_t).}$$

By the pairing identity $D_A(\boldsymbol{\eta}_0\|\boldsymbol{\eta}) = D_{A^*}(\boldsymbol{\eta}^*\|\boldsymbol{\eta}_0^*)$, this is equivalently

$$\boxed{\psi_{t+1} = \arg\min_{\psi\in\Psi} D_{A^*}(\tilde{\boldsymbol{\eta}}_t^*\|\boldsymbol{\eta}^*(\psi)).}$$

Hence curved mirror gradient descent consists of:

1. taking the unconstrained ambient mirror step along the m-geodesic in dual coordinates,
2. projecting back to the curved model by Bregman projection.

If $\boldsymbol{\eta}(\Psi)$ is e-flat, then this is exactly the e-projection of $\tilde{\boldsymbol{\eta}}_t$ onto $\boldsymbol{\eta}(\Psi)$, and the generalized Pythagorean theorem gives

$$D_A(\boldsymbol{\eta}(\psi)\|\tilde{\boldsymbol{\eta}}_t) = D_A(\boldsymbol{\eta}(\psi)\|\boldsymbol{\eta}(\psi_{t+1})) + D_A(\boldsymbol{\eta}(\psi_{t+1})\|\tilde{\boldsymbol{\eta}}_t) \qquad \text{for all } \psi \in \Psi.$$

If $\boldsymbol{\eta}^*(\Psi)$ is m-flat, the dual projection is governed by the dual Pythagorean identity in the same way.

For a genuinely curved family one should not expect a global Pythagorean theorem in general; what remains is the constrained minimization problem and its local orthogonality condition.

Orthogonality and the three-point identity. At the minimizer, first-order optimality gives

$$\langle \tilde{\boldsymbol{\eta}}_t^* - \boldsymbol{\eta}^*(\psi_{t+1}),\, \boldsymbol{v} \rangle = 0 \qquad \text{for all } \boldsymbol{v} \in T_{\boldsymbol{\eta}(\psi_{t+1})}\boldsymbol{\eta}(\Psi).$$

This is exactly the orthogonality term in the three-point identity. Thus the residual m-direction and every tangent e-direction at $\boldsymbol{\eta}(\psi_{t+1})$ meet orthogonally. If the embedded model is flat, this local statement upgrades to the full projection theorem and the corresponding Pythagorean identity.

1.9 The six NEF–QVF families: divergences and mirror updates

We now specialize to the six one-dimensional NEF–QVF families. Write

$$\mu := \eta^*(\eta) = A'(\eta), \qquad V(\mu) = A''(\eta(\mu)),$$

and let A^* be the Legendre transform of A. The primal and dual Bregman divergences are

$$D_A(\eta_1 \| \eta_0) = A(\eta_1) - A(\eta_0) - \mu_0(\eta_1 - \eta_0),$$

$$D_{A^*}(\mu_1 \| \mu_0) = A^*(\mu_1) - A^*(\mu_0) - \eta_0(\mu_1 - \mu_0), \qquad \eta_i = A^{*\prime}(\mu_i) = \eta(\mu_i),$$

and satisfy the Legendre–Bregman pairing

$$D_A(\eta_1 \| \eta_0) = D_{A^*}(\mu_0 \| \mu_1).$$

To obtain explicit scalar updates, it is convenient to write mirror descent in the mean coordinate μ with mirror potential A^*. Let

$$g_t := \frac{\partial L}{\partial \mu}(\mu_t), \qquad h_t := \alpha_t\, g_t.$$

Then the mirror step is

$$\eta_{t+1} = \eta_t - h_t, \qquad \mu_{t+1} = \eta^*(\eta_t - h_t) = \eta^*\big(\eta(\mu_t) - h_t\big),$$

or equivalently

$$\eta(\mu_{t+1}) = \eta(\mu_t) - h_t.$$

1. Normal family. Let

$$A(\eta) = \frac{\sigma_0^2 \eta^2}{2}, \qquad \mu = \sigma_0^2 \eta, \qquad V(\mu) = \sigma_0^2.$$

Then

$$A^*(\mu) = \frac{\mu^2}{2\sigma_0^2}, \qquad \eta(\mu) = \frac{\mu}{\sigma_0^2}.$$

The divergences are

$$D_A(\eta_1 \| \eta_0) = \frac{(\mu_1 - \mu_0)^2}{2\sigma_0^2}, \qquad D_{A^*}(\mu_1 \| \mu_0) = \frac{(\mu_1 - \mu_0)^2}{2\sigma_0^2}.$$

The mirror update is

$$\mu_{t+1} = \mu_t - \sigma_0^2\, h_t.$$

2. Poisson family. Let

$$A(\eta) = e^{\eta}, \qquad \mu = e^{\eta}, \qquad V(\mu) = \mu.$$

Then

$$A^*(\mu) = \mu \log \mu - \mu, \qquad \eta(\mu) = \log \mu.$$

Hence

$$D_A(\eta_1 \| \eta_0) = \mu_1 - \mu_0 - \mu_0 \log \frac{\mu_1}{\mu_0},$$

$$D_{A^*}(\mu_1 \| \mu_0) = \mu_1 \log \frac{\mu_1}{\mu_0} - (\mu_1 - \mu_0).$$

The mirror update is

$$\mu_{t+1} = \mu_t \, e^{-h_t}.$$

3. Gamma family. Fix $r > 0$ and let

$$A(\eta) = -r \log(-\eta) + \log \Gamma(r), \qquad \eta < 0,$$

so that

$$\mu = \frac{r}{-\eta}, \qquad V(\mu) = \frac{\mu^2}{r}.$$

Up to an irrelevant additive constant,

$$A^*(\mu) = -r \log \mu, \qquad \eta(\mu) = -\frac{r}{\mu}.$$

Therefore

$$D_A(\eta_1 \| \eta_0) = r \left(\frac{\mu_0}{\mu_1} - 1 - \log \frac{\mu_0}{\mu_1} \right),$$

$$D_{A^*}(\mu_1 \| \mu_0) = r \left(\frac{\mu_1}{\mu_0} - 1 - \log \frac{\mu_1}{\mu_0} \right).$$

The mirror update is

$$\mu_{t+1} = \frac{r \, \mu_t}{r + h_t \mu_t}.$$

4. Binomial family. Fix $N \in \mathbb{N}$ and let

$$A(\eta) = N \log(1 + e^{\eta}), \qquad \mu = N \frac{e^{\eta}}{1 + e^{\eta}}, \qquad V(\mu) = \mu \left(1 - \frac{\mu}{N} \right).$$

Then, up to an additive constant,

$$A^*(\mu) = \mu \log \frac{\mu}{N} + (N - \mu) \log \frac{N - \mu}{N},$$

$$\eta(\mu) = \log \frac{\mu}{N - \mu}.$$

Hence

$$D_A(\eta_1 \| \eta_0) = \mu_0 \log \frac{\mu_0}{\mu_1} + (N - \mu_0) \log \frac{N - \mu_0}{N - \mu_1},$$

$$D_{A^*}(\mu_1 \| \mu_0) = \mu_1 \log \frac{\mu_1}{\mu_0} + (N - \mu_1) \log \frac{N - \mu_1}{N - \mu_0}.$$

The mirror update is

$$\mu_{t+1} = N \frac{\mu_t e^{-h_t}}{N - \mu_t + \mu_t e^{-h_t}}.$$

5. Negative binomial family. Fix $r > 0$ and let

$$A(\eta) = -r \log(1 - e^{\eta}), \qquad \eta < 0,$$

so that

$$\mu = \frac{re^{\eta}}{1 - e^{\eta}}, \qquad V(\mu) = \mu\left(1 + \frac{\mu}{r}\right).$$

Up to an additive constant,

$$A^*(\mu) = \mu \log \mu - (r + \mu) \log(r + \mu),$$

$$\eta(\mu) = \log \frac{\mu}{r + \mu}.$$

Hence

$$D_A(\eta_1 \| \eta_0) = \mu_0 \log \frac{\mu_0}{\mu_1} - (r + \mu_0) \log \frac{r + \mu_0}{r + \mu_1},$$

$$D_{A^*}(\mu_1 \| \mu_0) = \mu_1 \log \frac{\mu_1}{\mu_0} - (r + \mu_1) \log \frac{r + \mu_1}{r + \mu_0}.$$

The mirror update is

$$\mu_{t+1} = \frac{r\,\mu_t e^{-h_t}}{r + \mu_t - \mu_t e^{-h_t}}.$$

6. Generalized hyperbolic secant family. In the canonical form,

$$A(\eta) = -2r \log\big(2 \cos(\eta/2)\big), \qquad \eta \in (-\pi, \pi),$$

with

$$\mu = r \tan(\eta/2), \qquad V(\mu) = \frac{r}{2} + \frac{\mu^2}{2r}.$$

Then

$$\eta(\mu) = 2 \arctan(\mu/r),$$

and, up to an additive constant,

$$A^*(\mu) = 2\mu \arctan(\mu/r) - r \log(r^2 + \mu^2).$$

The primal divergence is

$$D_A(\eta_1 \| \eta_0) = r \log \frac{r^2 + \mu_1^2}{r^2 + \mu_0^2} - 2\mu_0 \left[\arctan\left(\frac{\mu_1}{r}\right) - \arctan\left(\frac{\mu_0}{r}\right)\right],$$

and the dual divergence is

$$D_{A^*}(\mu_1 \| \mu_0) = -r \log \frac{r^2 + \mu_1^2}{r^2 + \mu_0^2} + 2\mu_1 \left[\arctan\left(\frac{\mu_1}{r}\right) - \arctan\left(\frac{\mu_0}{r}\right)\right].$$

The mirror update is

$$\mu_{t+1} = r \tan\left(\arctan\frac{\mu_t}{r} - \frac{h_t}{2}\right).$$

Family	$\eta(\mu)$	mirror update μ_{t+1}
Normal	μ/σ_0^2	$\mu_t - \sigma_0^2 h_t$
Poisson	$\log \mu$	$\mu_t e^{-h_t}$
Gamma	$-r/\mu$	$\dfrac{r\,\mu_t}{r + h_t \mu_t}$
Binomial	$\log \frac{\mu}{N-\mu}$	$N \dfrac{\mu_t e^{-h_t}}{N - \mu_t + \mu_t e^{-h_t}}$
Negative binomial	$\log \frac{\mu}{r+\mu}$	$\dfrac{r\,\mu_t e^{-h_t}}{r + \mu_t - \mu_t e^{-h_t}}$
GHS	$2 \arctan(\mu/r)$	$r \tan\left(\arctan \dfrac{\mu_t}{r} - \dfrac{h_t}{2}\right)$

Table 6: Mean-coordinate mirror updates for the six NEF–QVF families, with $h_t = \alpha_t g_t$, $g_t = \partial L / \partial \mu(\mu_t)$, and $\eta(\mu_{t+1}) = \eta(\mu_t) - h_t$.