

## A Motivating our continuous-time objective function

In this appendix, we describe in detail how the most general form of our objective function (Eq. 1) can be viewed as a direct generalization of the objective we used to justify gradient descent (see Sec. 2).

### A.1 From a single-step objective to a multi-step objective

Recall that the objective we used to justify gradient descent (and Newton’s method) had the form

$$J(\Delta\theta) = \frac{\|\Delta\theta\|^2}{2\eta} + \mathcal{L}(\theta + \Delta\theta), \quad (17)$$

where we have neglected to Taylor expand  $\mathcal{L}$  to keep our objective more general. Note that  $J$  involves two terms: one which penalizes large parameter changes, and another which penalizes high values of the loss. The former term is especially necessary if we consider a local approximation (e.g., an expansion in powers of  $\Delta\theta$ ) of  $\mathcal{L}$ , since the approximation may no longer be valid if we consider sufficiently large steps.

We would like to go from this single-step objective to an analogous multi-step objective. The most obvious way to do this is to define the objective over a sum of  $K$  terms, each of which involves a parameter change penalty and a loss term:

$$J_{\text{multi}}(\Delta\theta_0, \Delta\theta_1, \dots, \Delta\theta_{K-1}) := \sum_{t=0}^{K-1} \frac{1}{2\eta} \|\Delta\theta_t\|^2 + \mathcal{L}(\theta_t). \quad (18)$$

Note the philosophy of including the loss at each step: it implies that we would like a path through parameter space that involves decreases to the loss at each step, rather than just at the end. It is also possible to penalize the loss only at the end, but the resulting learning rules would look somewhat different than, e.g., variants of gradient descent.

### A.2 From discrete time to continuous time

Although we could directly study the optimization of Eq. 18, we can instead exploit the fact that objectives like this tend to be easier to analyze in continuous time, since determining the optimal sequence  $\Delta\theta_0, \Delta\theta_1, \dots$  becomes a well-studied calculus of variations [28, 30] problem. If each step takes an amount of ‘time’  $\Delta t$ , the cost of each step is scaled to be proportional to  $\Delta t$ , and we adjust  $\eta \rightarrow \eta(\Delta t)^2$  for dimensional reasons, we obtain a continuous-time objective that directly generalizes Eq. 18:

$$J(\{\theta_t\}) = \sum_{t=0}^{K-1} \left[ \frac{1}{2\eta} \left\| \frac{\Delta\theta_t}{\Delta t} \right\|^2 + \mathcal{L}(\theta_t) \right] \Delta t \xrightarrow{\Delta t \rightarrow 0} \int_0^T \frac{1}{2\eta} \|\dot{\theta}_t\|^2 + \mathcal{L}(\theta_t) dt \quad (19)$$

where  $T$  is the total optimization ‘time’. Note that our optimization problem now looks like an optimal control problem (with a control cost  $\propto 1/\eta$  and state cost  $\mathcal{L}$ ) or classical mechanics problem (where a particle with mass  $m := 1/\eta$  moves in a potential determined by  $\mathcal{L}$ ).

To make it slightly clearer how our solution depends on  $\mathcal{L}$ , and to more explicitly control the influence of the loss, we will add a prefactor  $k > 0$  (which has units of inverse time) in front of it:

$$J(\{\theta_t\}) = \int_0^T \frac{1}{2\eta} \|\dot{\theta}_t\|^2 + k\mathcal{L}(\theta_t) dt. \quad (20)$$

This change does not meaningfully affect our optimization problem; it only amounts to a change in the units of  $J$ .

### A.3 Incorporating temporal discounting

Our objective (Eq. 21) now has the form of an optimal control problem. Motivated by this observation, we incorporate a temporal discounting factor  $e^{-\gamma t}$ , which causes the learner to overemphasize rewards and costs that are nearer in time:

$$J(\{\theta_t\}) = \int_0^T \left[ \frac{1}{2\eta} \|\dot{\theta}_t\|^2 + k\mathcal{L}(\theta_t) \right] e^{-\gamma t} dt. \quad (21)$$



We could have included a more general discounting factor, as some prior work does [52]. We restrict ourselves to exponential discounting because it is a canonical choice [26, 27] and yields relatively simple EL equations. As an aside, allowing a more general discounting factor would allow us to derive learning rules that involve Nesterov momentum rather than just standard (Polyak) momentum.

Since we are not generally interested in any specific learning time  $T$ , and since taking the  $T \rightarrow \infty$  substantially simplifies some EL-equation-related math, we will consider the  $T \rightarrow \infty$  version of the objective in everything that follows:

$$J(\{\theta_t\}) = \int_0^\infty \left[ \frac{1}{2\eta} \|\dot{\theta}_t\|^2 + k\mathcal{L}(\theta_t) \right] e^{-\gamma t} dt. \quad (22)$$

This objective is *infinitely* forward-looking, in the sense that trajectories which optimize it depend on considering  $\theta_t$  at arbitrary distant future times.

#### A.4 Incorporating parameter space geometry

Our original single-step objective (Eq. 17) implicitly assumes that parameter space is Euclidean, or at least that a Euclidean distance metric is most appropriate for penalizing large parameter changes. As researchers like Amari [21] have observed, this is not necessarily true. More generally, we might want to penalize distances according to a less trivial metric like the Fisher information metric.

We can account for this fact by modifying Eq. 17 to involve a metric  $\mathbf{G}$ :

$$J(\Delta\theta) = \frac{1}{2\eta} (\Delta\theta)^T \mathbf{G}(\theta) (\Delta\theta) + \mathcal{L}(\theta + \Delta\theta). \quad (23)$$

Note that, for all  $\theta$ , we will assume  $\mathbf{G}(\theta)$  is symmetric and positive definite (and hence invertible). To account for partial controllability (see Sec. 4), we could make the similarly minor change of penalizing not the size of  $\Delta\theta$ , but the size of  $\Delta\theta - \mathbf{f}(\theta)$  (i.e., the size of deviations from the ‘default’ dynamics determined by the drift function  $\mathbf{f}$ ):

$$J(\Delta\theta) = \frac{1}{2\eta} [\Delta\theta - \mathbf{f}(\theta)]^T \mathbf{G}(\theta) [\Delta\theta - \mathbf{f}(\theta)] + \mathcal{L}(\theta + \Delta\theta). \quad (24)$$

We can generalize this objective to something which operates in continuous-time over multiple steps by the same argument as behavior. We must only make the small change  $\mathbf{f} \rightarrow \mathbf{f}\Delta t$  so that the continuous-time limit is well-defined. We obtain

$$J(\{\theta_t\}) = \int_0^\infty \left[ \frac{1}{2\eta} [\dot{\theta}_t - \mathbf{f}(\theta_t)]^T \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] + k\mathcal{L}(\theta_t) \right] e^{-\gamma t} dt. \quad (25)$$

#### A.5 Accounting for loss approximation

Lastly, we must account for our assumption that the true loss  $\mathcal{L}$  is generally only partially observable. Since we have already posed learning as a reinforcement learning and optimal control problem, this change is easy to make. In those settings, one accounts for random variables by averaging the objective over them; we will do the same here. We finally have

$$J(\{\theta_t\}) = \mathbb{E}_{\{\hat{\mathcal{L}}_t\}} \left\{ \int_0^\infty \left( \frac{1}{2\eta} [\dot{\theta}_t - \mathbf{f}(\theta_t)]^T \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] + k\hat{\mathcal{L}}_t(\theta_t) \right) e^{-\gamma t} dt \right\}. \quad (26)$$



## B Experiment details

In this appendix, we provide details relevant to understanding the numerical experiments mentioned in the main text. See <https://github.com/john-vastola/lossnav-neurips25> for code that reproduces all figures.

**Direct optimization of the objective.** In Fig. 1, we numerically estimate the minimizer of the objective

$$J(\{\theta_t\}) = \int_0^\infty \left( \frac{1}{2\eta} \|\dot{\theta}_t\|^2 + k\mathcal{L}(\theta_t) \right) e^{-\gamma t} dt = \int_0^\infty \left( \frac{\dot{\theta}_1^2}{2\eta} + \frac{\dot{\theta}_2^2}{2\eta} + k\mathcal{L}(\theta_1, \theta_2) \right) e^{-\gamma t} dt \quad (27)$$

given a double-well loss

$$\mathcal{L}(\theta_1, \theta_2) = a(\theta_1^2 - b)^2 + c\theta_2^2 + d\theta_1 \quad (28)$$

with  $a = 1$ ,  $b = 1$ ,  $c = 1$ , and  $d = -1/2$ , assuming  $k = \eta = 1$  and  $\gamma = 0.1$ . We do this optimization directly (rather than via the EL equations) by discretizing  $\theta_1(t)$  and  $\theta_2(t)$ , i.e.,

$$J \approx \sum_{k=1}^N \left( \frac{[\theta_1(t_{k+1}) - \theta_1(t_k)]^2}{2\eta(\Delta t)^2} + \frac{[\theta_2(t_{k+1}) - \theta_2(t_k)]^2}{2\eta(\Delta t)^2} + k\mathcal{L}(\theta_1(t_k), \theta_2(t_k)) \right) e^{-\gamma t_k} (\Delta t) \quad (29)$$

where we choose  $N + 1$  equally spaced time points  $t_0, t_1, \dots, t_N$  for simplicity. This means that  $t_k := k\Delta t$  for all  $k = 0, \dots, N$ , where  $\Delta t := T/N$ . Here, the cutoff time  $T > 0$  is chosen to be large enough that the optimal trajectory is near the global minimum of the loss at the final time point  $t_N$ . (This means that, even though we do not integrate all the way until  $t = \infty$ , not much interesting happens beyond time  $T$ .)

By fixing the initial point  $\theta(t_0) = (\theta_1(t_0), \theta_2(t_0))^T$ , final point  $\theta(t_N) = (\theta_1(t_N), \theta_2(t_N))^T$ , and the cutoff time  $T$ , we can vary the remaining  $2(N - 1)$  degrees of freedom to determine the optimal trajectory. Following Strang et al. [29], we minimize  $J$  with respect to these variables using a PyTorch-based gradient descent approach.

Given the solution, we can estimate the ‘kinetic energy’ throughout a trajectory by computing

$$\text{KE}(t_k) := \left( \frac{[\theta_1(t_{k+1}) - \theta_1(t_k)]^2}{2\eta(\Delta t)^2} + \frac{[\theta_2(t_{k+1}) - \theta_2(t_k)]^2}{2\eta(\Delta t)^2} \right) \quad (30)$$

and the ‘potential energy’ by computing

$$\text{PE}(t_k) := k\mathcal{L}(\theta_1(t_k), \theta_2(t_k)) . \quad (31)$$

**Application of ballistic rule to MNIST and CIFAR-10 image classification.** For Fig. 2, we implement the ‘ballistic’ rule that emerges from one of our exact solutions in the  $\gamma = 0$  limit (see Eq. 7), and use it to train classifiers on the MNIST [55] and CIFAR-10 [56] image datasets. At each step, the ballistic rule prescribes a parameter update proportional to

$$\Delta\theta \propto -H^{-1/2}g \quad (32)$$

where  $g$  is the current loss gradient and  $H$  is the current Hessian of the loss. Since Hessian computation is expensive and difficult to scale, we implement the ballistic rule by using an Adam-like [13] approach: at each step, we update a running average of (uncentered) gradient variances along each direction. This involves three crude approximations: we use this running average instead of directly computing the Hessian; we only compute the diagonal entries of the Hessian proxy; and we do not center the variance estimates. Despite these approximations, we still believe that this heuristic approach captures the spirit of the ballistic rule. Furthermore, using an Adam-like approach is theoretically reasonable given the link we identify between Adam and our ballistic rule (Sec. 5).

Let  $v_i$  denote the running gradient variance associated with the parameter  $\theta_i$ . For each  $i$ , the precise updates per step are

$$\begin{aligned} v_i &\leftarrow \beta_2 v_i + (1 - \beta_2) g_i^2 \\ \Delta\theta_i &= -\eta g_i / \sqrt{v_i} \end{aligned} \quad (33)$$



where  $g_i$  denotes the current gradient along the  $\theta_i$  direction,  $\eta$  denotes the learning rate, and  $\beta_2 \in [0, 1]$  controls the time scale on which gradient observations are averaged. Note that this usage of  $\beta_2$  is intended to match Adam’s; like in Adam, values like  $\beta_2 = 0.9$  and  $\beta_2 = 0.999$  (i.e., values close to one) appear to work well.

We consider only two simple architectures to illustrate the ballistic rule’s behavior:

- a multilayer perceptron (MLP) with two hidden layers, which we trained on MNIST; and
- a small convolutional neural network (CNN) with four convolutional layers, two max pooling layers, and a final fully-connected layer, which we trained on CIFAR-10.

We generally found that, especially since this implementation of the ballistic rule is similar to the standard implementation of Adam, hyperparameter settings that work well for Adam also work well for it. For example, learning rates around  $1e-3$  and  $1e-4$  worked well. For MNIST,  $\beta_2 = 0.999$  performed reasonably well, but for CIFAR-10 a lower value ( $\beta_2 = 0.9$ ) appeared to be necessary for good performance.



## C Deriving learning rules via the Euler-Lagrange equations

Suppose that the parameter vector  $\theta$  is  $D$ -dimensional. After averaging over the loss landscape belief model in Eq. 1, our objective has the form

$$J(\{\theta_t\}) = \int_0^\infty \left[ \frac{1}{2\eta} [\dot{\theta}_t - \mathbf{f}(\theta_t)]^T \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] + k\mathcal{L}(\theta_t) \right] e^{-\gamma t} dt \quad (34)$$

for some  $\mathcal{L}$ ,  $\mathbf{G}$ , and so on.<sup>4</sup> Given an objective like this, how do we go about (analytically) deriving learning rules?

We are looking for a trajectory  $\{\theta_t\}_{t \in [0, \infty)}$  which minimizes  $J$ . This trajectory is not necessarily unique, for example due to symmetry, but is often unique in practice. By the calculus of variations, subject to the relevant boundary conditions (and smoothness-related technical conditions), the optimal trajectory can be shown [28, 30] to satisfy the *Euler-Lagrange (EL) equations*.

The idea is to use the objective to define a Lagrangian

$$L(\theta, \dot{\theta}, t) := \left[ \frac{1}{2\eta} [\dot{\theta} - \mathbf{f}(\theta)]^T \mathbf{G}(\theta) [\dot{\theta} - \mathbf{f}(\theta)] + k\mathcal{L}(\theta) \right] e^{-\gamma t}. \quad (35)$$

The Euler-Lagrange equations are the  $D$  equations

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}} \right) = \frac{\partial L}{\partial \theta} \quad , \text{ i.e., } \quad \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}_i} \right) = \frac{\partial L}{\partial \theta_i} \quad \text{for } i = 1, \dots, D. \quad (36)$$

More explicitly, we have  $D$  second-order ODEs

$$\frac{d}{dt} \left( \mathbf{G}(\theta) [\dot{\theta} - \mathbf{f}(\theta)] \right) - \gamma \mathbf{G}(\theta) [\dot{\theta} - \mathbf{f}(\theta)] = \eta k \nabla_{\theta} \mathcal{L}(\theta) + \frac{\partial}{\partial \theta} \left( \frac{1}{2} [\dot{\theta} - \mathbf{f}(\theta)]^T \mathbf{G}(\theta) [\dot{\theta} - \mathbf{f}(\theta)] \right). \quad (37)$$

Since they are second-order, **specifying solutions to these ODEs** (up to symmetry-related degeneracies) **requires two pieces of boundary data**. In classical mechanics [28], one often looks for solutions with a specified *initial position* and *initial momentum* (or equivalently, initial velocity). In our context, this does not quite make sense. We *do* want a solution with a specified initial parameter vector  $\theta_0$ , but requiring some particular initial velocity or momentum is less obviously meaningful.

The second piece of boundary data follows from the fact that we would like  $J$  to be *minimized*. This is equivalent to requiring that the asymptotic ( $t \rightarrow \infty$ ) endpoint of the trajectory corresponds to the global minimum of the loss, at least if  $\mathbf{f} \equiv \mathbf{0}$ . If  $\mathbf{f}$  is nonzero, the drift contributes a term to the ‘effective’ loss, and it is the global minimum of *this* function which must be asymptotically approached instead.

**In short, we can find optimal trajectories  $\{\theta_t\}_{t \in [0, \infty)}$  by solving the EL equations (Eq. 37) subject to the constraints that:**

- $\theta_0$  is fixed, and corresponds to the ‘current’ parameter vector.
- The remaining degree of freedom (e.g., the trajectory endpoint or initial momentum) is chosen so that  $J$  is minimized.

Below, we provide two simple but instructive one-dimensional examples that provide intuition about how this process plays out in practice.

### C.1 Example: quadratic loss

Suppose  $D = 1$ ,  $\mathbf{G} = \mathbf{I}$ ,  $\mathbf{f} \equiv \mathbf{0}$ , and

$$\mathcal{L}(\theta) = \frac{h}{2} (\theta - \theta_*)^2, \quad (38)$$

<sup>4</sup>If the loss model  $\hat{\mathcal{L}}$  involves parameters (e.g., estimated gradients and curvature) which are themselves dynamic, these parameters should be considered components of  $\theta$ , and can hence contribute terms to an ‘effective’ metric  $\mathbf{G}$ , an effective drift  $\mathbf{f}$ , etc. This is easiest to see in the context of specific examples, like the ones that appear in Sec. 5.



i.e., the loss is quadratic with a global minimum at  $\theta_*$ . Eq. 34 becomes

$$J(\{\theta_t\}) = \int_0^\infty \left[ \frac{1}{2\eta} \dot{\theta}_t^2 + \frac{kh}{2} (\theta_t - \theta_*)^2 \right] e^{-\gamma t} dt. \quad (39)$$

The corresponding Lagrangian is

$$L(\theta, \dot{\theta}, t) = \left[ \frac{1}{2\eta} \dot{\theta}^2 + \frac{kh}{2} (\theta - \theta_*)^2 \right] e^{-\gamma t}, \quad (40)$$

and the corresponding EL equation is

$$\ddot{\theta}_t - \gamma \dot{\theta}_t = \eta kh (\theta_t - \theta_*) \implies \ddot{\theta}_t - \gamma \dot{\theta}_t - \eta kh \theta_t = -\eta kh \theta_*. \quad (41)$$

This is a linear, second-order ODE with constant coefficients, and so can be solved in the usual way<sup>5</sup>. This ODE's characteristic equation is

$$r^2 - \gamma r - \eta kh = 0 \quad (42)$$

and has roots

$$r_\pm = \frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} + \eta kh}. \quad (43)$$

Importantly, one of these roots is positive and one is negative, a fact which we will return to shortly. The full solution can hence be written

$$\theta_t = \theta_* + c_+ e^{r_+ t} + c_- e^{r_- t} \quad (44)$$

where  $\theta_*$  is the (obvious) particular solution. Enforcing the initial condition,

$$\theta_0 = \theta_* + c_+ + c_- \implies c_- = \theta_0 - \theta_* - c_+. \quad (45)$$

Hence,

$$\begin{aligned} \theta_t &= \theta_* + c_+ e^{r_+ t} + (\theta_0 - \theta_* - c_+) e^{r_- t} \\ \dot{\theta}_t &= c_+ r_+ e^{r_+ t} + (\theta_0 - \theta_* - c_+) r_- e^{r_- t}. \end{aligned} \quad (46)$$

We can determine  $c_+$  by substituting these into  $J$  and minimizing it with respect to  $c_+$ . Note,

$$\begin{aligned} (\theta_t - \theta_*)^2 &= c_+^2 e^{2r_+ t} + (\theta_0 - \theta_* - c_+)^2 e^{2r_- t} + 2c_+(\theta_0 - \theta_* - c_+) e^{(r_+ + r_-)t} \\ \dot{\theta}_t^2 &= c_+^2 r_+^2 e^{2r_+ t} + (\theta_0 - \theta_* - c_+)^2 r_-^2 e^{2r_- t} + 2c_+ r_+ r_- (\theta_0 - \theta_* - c_+) e^{(r_+ + r_-)t}. \end{aligned} \quad (47)$$

Since

$$\begin{aligned} r_+^2 &= \gamma r_+ + \eta kh \\ r_-^2 &= \gamma r_- + \eta kh \\ r_+ r_- &= -\eta kh, \end{aligned} \quad (48)$$

the  $\dot{\theta}_t^2$  expression can be simplified to

$$\begin{aligned} \dot{\theta}_t^2 &= \eta kh \left[ c_+^2 e^{2r_+ t} + (\theta_0 - \theta_* - c_+)^2 e^{2r_- t} - 2c_+(\theta_0 - \theta_* - c_+) e^{(r_+ + r_-)t} \right] \\ &\quad + \gamma \left[ c_+^2 r_+ e^{2r_+ t} + (\theta_0 - \theta_* - c_+)^2 r_- e^{2r_- t} \right]. \end{aligned} \quad (49)$$

After some algebra, the integrand of  $J$  is

$$\left( kh + \frac{\gamma}{2\eta} r_+ \right) c_+^2 e^{(2r_+ - \gamma)t} + \left( kh + \frac{\gamma}{2\eta} r_- \right) (\theta_0 - \theta_* - c_+)^2 e^{(2r_- - \gamma)t}. \quad (50)$$

Integrating from  $t = 0$  to  $t = T$ ,  $J$  equals

$$J = \lim_{T \rightarrow \infty} \left( kh + \frac{\gamma}{2\eta} r_+ \right) c_+^2 \frac{(e^{(2r_+ - \gamma)T} - 1)}{2r_+ - \gamma} + \left( kh + \frac{\gamma}{2\eta} r_- \right) (\theta_0 - \theta_* - c_+)^2 \frac{(e^{(2r_- - \gamma)T} - 1)}{2r_- - \gamma}. \quad (51)$$

<sup>5</sup>Physically, it is analogous to a damped harmonic oscillator, but with the ‘wrong’ sign on the  $\theta_t$  coefficient. This difference makes sense, since a learning trajectory ought to eventually settle down into a global minimum rather than oscillate.



Recall that  $r_+$  is positive. The quantity

$$2r_+ - \gamma = \sqrt{\frac{\gamma^2}{4} + \eta k h} \quad (52)$$

is also positive, which means that the term with  $e^{(2r_+ - \gamma)T}$  approaches infinity in the  $T \rightarrow \infty$  limit we're interested in. Inspecting Eq. 51, the only way to remove the offending term is to set  $c_+ = 0$ .

But this outcome was foreseeable if we note that the positive-root term  $e^{r_+ t}$  'blows up' in the long-time limit, whereas the other term doesn't. For this reason, in future derivations we will bypass this argument and simply set  $c_+$  (or its higher-dimensional analogue) to zero.

Incidentally, if our objective involved a finite time horizon  $t \in [0, T]$  rather than  $t \in [0, \infty)$ , in general we would have  $c_+ \neq 0$ , which would somewhat complicate our expressions for optimal learning trajectories.

Setting  $c_+ = 0$ , we finally find that the optimal learning trajectory has

$$\begin{aligned} \theta_t &= \theta_* + (\theta_0 - \theta_*)e^{r_- t} \\ \dot{\theta}_t &= (\theta_0 - \theta_*)r_- e^{r_- t}, \end{aligned} \quad (53)$$

i.e., it approaches the global minimum exponentially quickly, at a rate  $r_-$ .

## C.2 Example: double-well loss

While the previous example is instructive, the loss function we considered was convex, and involved only one (local/global) minimum. It is somewhat more interesting to see what happens with a double-well loss

$$\mathcal{L}(\theta) = \frac{h}{4}(\theta^2 - \theta_*^2)^2 - \frac{q}{3}\theta^3 - \mathcal{L}_{min} \quad (54)$$

where  $h > 0$  and  $q > 0$ , and where the additive constant  $\mathcal{L}_{min}$  is chosen so that  $\mathcal{L}$  equals zero at its global minimum. This loss has two local minima, as is clear from its derivative:

$$\mathcal{L}'(\theta) = h\theta(\theta^2 - \theta_*^2) - q\theta^2. \quad (55)$$

In particular, since

$$\mathcal{L}'(\theta) = 0 \implies h\theta \left[ \theta^2 - \frac{q}{h}\theta - \theta_*^2 \right], \quad (56)$$

the two minima are at

$$\theta_{\pm} := \pm \sqrt{\theta_*^2 + \frac{q^2}{4h^2}} + \frac{q}{2h}, \quad (57)$$

and are near  $\pm\theta_*$  if  $q$  is small. The lower minimum is at  $\theta_+$ ; the  $h$  term is identical at both  $\theta_+$  and  $\theta_-$ , but the  $q$  term is negative at  $\theta_+$ .

For this loss, the objective  $J$  is

$$J(\{\theta_t\}) = \int_0^\infty \left[ \frac{1}{2\eta} \dot{\theta}_t^2 + k \left( \frac{h}{4}(\theta_t^2 - \theta_*^2)^2 - \frac{q}{3}\theta_t^3 \right) \right] e^{-\gamma t} dt \quad (58)$$

and the EL equation is

$$\ddot{\theta}_t - \gamma \dot{\theta}_t = \eta k \mathcal{L}'(\theta_t) = \eta k h \theta_t (\theta_t - \theta_+) (\theta_t - \theta_-). \quad (59)$$

This second-order ODE is highly nonlinear, and it is not obvious if it is analytically solvable. If we assume  $\gamma = 0$  (i.e., no temporal discounting), the EL equation becomes

$$\ddot{\theta}_t = \eta k \mathcal{L}'(\theta_t). \quad (60)$$

This can be simplified somewhat by using a trick. If we multiply the left-hand side and right-hand side by  $\dot{\theta}_t$ , we can integrate both with respect to time; we obtain

$$\dot{\theta}_t^2 = 2\eta k \mathcal{L}(\theta_t) + 2\eta E \quad (61)$$

where  $E$  is a constant. Since the left-hand side is nonnegative,  $E \geq -k\mathcal{L}(\theta_{min})$ , where  $\theta_{min}$  is the argument for which  $\mathcal{L}$  achieves its global minimum. (Here, due to the additive offset we included,  $\mathcal{L}(\theta_{min}) = 0$ .)



We label this constant  $E$  since it corresponds to this setting's notion of energy, as we could figure out from an analysis of this problem's Hamiltonian<sup>6</sup>.

Eq. 61 implies that

$$J = \lim_{T \rightarrow \infty} \int_0^T \frac{1}{2\eta} \dot{\theta}_t^2 + k\mathcal{L}(\theta_t) dt = \lim_{T \rightarrow \infty} \int_0^T 2k\mathcal{L}(\theta_t) + E dt = \lim_{T \rightarrow \infty} \int_0^T 2k\mathcal{L}(\theta_t) dt + ET \quad (62)$$

along the optimal trajectory. But this is a problem, since  $ET \rightarrow \infty$  as  $T \rightarrow \infty$  for most choices of  $E$ . We do best by setting energy equal to its minimum possible value  $E = -k\mathcal{L}(\theta_{min}) = 0$ , since the problematic term vanishes and the loss term asymptotically approaches zero (since the value of the loss at the global minimum is zero). That is,

$$J = \lim_{T \rightarrow \infty} \int_0^T 2k\mathcal{L}(\theta_t) dt < \infty. \quad (63)$$

Going back to Eq. 61, this means that

$$\dot{\theta}_t = \pm \sqrt{2\eta k\mathcal{L}(\theta_t)}. \quad (64)$$

This defines an optimal learning trajectory, and also an optimal learning rule.

What does this equation mean? To understand this expression, suppose  $\theta_0 = \theta_-$ , i.e., that the learner begins in the shallower minimum. Clearly, the optimal learning dynamics must move right (towards  $\theta_+$ ); the optimal trajectory in this case has

$$\dot{\theta}_t = \sqrt{2\eta k\mathcal{L}(\theta_t)}, \quad (65)$$

i.e., we take the plus sign. If the learner begins at a parameter  $\theta > \theta_+$ , we would instead take the minus sign.

Although this example is somewhat complicated, there are two important takeaways:

- Long-horizon optimal learning trajectories approach the global minimum rather than any local minima.
- Even if the functional form of  $\mathcal{L}$  is more complicated than quadratic, it is in some cases possible to derive optimal learning rules and trajectories.

In the rest of this paper, partly because analyses like the above are difficult, and partly because our goal is to derive well-known learning rules, we essentially only consider quadratic (local) approximations of the loss.

---

<sup>6</sup>See Vastola [44] for the details of how to follow this line of thought in an analogous theoretical setting. At least if  $\gamma = 0$ , this notion of energy is conserved along the optimal trajectory.



## D More on the boundary conditions of the Euler-Lagrange equations

The boundary conditions for the objective minimization problem we describe in Sec. 2 are reasonably clear: of all the possible well-behaved (e.g., smooth, bounded) trajectories  $\{\theta_t\}_{t \in [0, \infty)}$  through the  $D$ -dimensional parameter space which begin at a prescribed initial parameter vector  $\theta_0$ , the ‘optimal’ trajectory is the one which makes  $J$  smallest. ‘Smallest’ here is a well-defined notion, since  $J$  is bounded from below as long as the loss  $\mathcal{L}$  is bounded from below. If multiple trajectories make  $J$  as small as possible, then each of them is optimal.

However, the boundary conditions for the Euler-Lagrange (EL) equations are less obvious. The EL equations answer the following question: given all possible smooth trajectories that begin at  $\theta_0$  and reach  $\theta(T)$  at time  $T > 0$ , which of them makes the Lagrangian (i.e., the integrand of the objective) *stationary*? This is three steps removed from our original minimization problem, since (i) stationary points may not correspond to local minima, (ii) local minima may not be global minima, and (iii) a given  $\theta(T)$  may not correspond to an optimal trajectory.

We think about the boundary conditions of the EL equations we encounter in this paper in the following way. First, we assume that the trajectory of interest has a prescribed initial point  $\theta_0$ , since this is a boundary condition of the original minimization problem. Second, since (according to standard calculus of variations results, under mild assumptions<sup>7</sup>) the Lagrangian  $L$  is stationary at the global minimizer of  $J$ , satisfying the EL equations is a *necessary* (but not sufficient) condition for a trajectory to be optimal. Third, any two trajectories can be compared (i.e., which corresponds to a lower value of  $J$ ?).

Together, the first two insights tell us that the EL equations must be satisfied for a specific initial point  $\theta_0$  and some other point  $\theta(T)$ . The third insight tells us that, if we do not know which  $\theta(T)$  to use, we can compare any two possibilities by comparing the corresponding values of  $J$ . This yields a two-level optimization strategy: for many endpoints  $\theta(T)$ , solve the EL equations; then, choose the solution whose corresponding  $J$  is smallest.

This strategy is not circular, and is useful because it reduces an optimization problem defined over an infinite-dimensional function space (i.e., the infinite-dimensional space of all possible parameter trajectories) to an optimization problem over a  $D$ -dimensional space (since each possible solution corresponds to a particular choice of  $\theta(T) \in \mathbb{R}^D$ ).

Finally, it is worth noting that the rather abstract conditions under which a calculus of variations problem is interesting and well-defined are not that useful for understanding many of the extremely simple objectives we consider in this paper. Especially given a quadratic loss, we can often verify directly (through algebra rather than numerics) that solutions of the EL equations correspond to global minimizers of  $J$ , and even directly compute  $J$  as a function of any  $\theta_0$  and  $\theta(T)$ .

---

<sup>7</sup>For example, we may want to assume that the kinetic term is convex and coercive. These two properties hold for the simple quadratic kinetic terms we consider throughout.



## E Deriving learning rules with momentum: more details

In this appendix, we derive the results presented in Sec. 3. The relevant objective is

$$J(\{\boldsymbol{\theta}_t\}) = \int_0^\infty \left( \frac{1}{2\eta} \|\dot{\boldsymbol{\theta}}_t\|^2 + k\mathcal{L}(\boldsymbol{\theta}_t) \right) e^{-\gamma t} dt. \quad (66)$$

The corresponding Lagrangian is

$$L(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}, t) = \left( \frac{1}{2\eta} \|\dot{\boldsymbol{\theta}}\|^2 + k\mathcal{L}(\boldsymbol{\theta}) \right) e^{-\gamma t} \quad (67)$$

and the corresponding EL equations are

$$\ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t = \eta k \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t). \quad (68)$$

We can rewrite these equations in a suggestive form by defining ‘momentum’ as  $\mathbf{p}_t := \dot{\boldsymbol{\theta}}_t$ , so that we have

$$\dot{\boldsymbol{\theta}}_t = \mathbf{p}_t \quad \dot{\mathbf{p}}_t = \gamma \mathbf{p}_t + \eta k \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t). \quad (69)$$

Note that our convention for momentum matches machine learning practice, but does not necessarily match the physics convention for momentum, which has  $\mathbf{p}_t := \frac{\partial L}{\partial \dot{\boldsymbol{\theta}}_t}$ .

### E.1 Solution to the Euler-Lagrange equations for a quadratic loss

If we assume a quadratic loss

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_t) := \mathcal{L}(\boldsymbol{\theta}_0) + \mathbf{g}^T (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^T \mathbf{H} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0), \quad (70)$$

where  $\mathbf{g}$  is the local gradient and  $\mathbf{H}$  is the (symmetric, positive definite) local Hessian, the EL equations (Eq. 68) become

$$\ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t = \eta k [\mathbf{g} + \mathbf{H}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)] \implies \ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t - \eta k \mathbf{H} \boldsymbol{\theta}_t = \eta k [\mathbf{g} - \mathbf{H} \boldsymbol{\theta}_0]. \quad (71)$$

The above represents a system of coupled second-order ODEs. We can decouple it by exploiting an eigendecomposition of  $\mathbf{H}$ . Since  $\mathbf{H} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T$  for some orthogonal matrix  $\mathbf{Q}$  and diagonal matrix  $\boldsymbol{\Lambda}$  (whose diagonal entries are nonnegative), we can premultiply both sides of this equation by  $\mathbf{Q}^T$  to obtain

$$\ddot{\boldsymbol{\phi}}_t - \gamma \dot{\boldsymbol{\phi}}_t - \eta k \boldsymbol{\Lambda} \boldsymbol{\phi}_t = \eta k [\tilde{\mathbf{g}} - \boldsymbol{\Lambda} \boldsymbol{\phi}_0], \quad (72)$$

where we define  $\boldsymbol{\phi}_t := \mathbf{Q}^T \boldsymbol{\theta}_t$  and  $\tilde{\mathbf{g}} := \mathbf{Q}^T \mathbf{g}$ . We now have many linear, second-order ODEs identical in form to the one from the first example in Appendix C. For a given  $\phi_i$ , we have

$$\ddot{\phi}_i - \gamma \dot{\phi}_i - \eta k \lambda_i \phi_i = \eta k [\tilde{g}_i - \lambda_i \phi_{i0}], \quad (73)$$

where  $\lambda_i := \Lambda_{ii} \geq 0$ . The corresponding characteristic equation reads

$$r^2 - \gamma r - \eta k \lambda_i = 0 \quad (74)$$

and has roots

$$r_{\pm} = \frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} + \eta k \lambda_i}. \quad (75)$$

By the same argument we used in Appendix C, we throw away the positive root (since it does not minimize  $J$ ) and keep the negative one. Denote the (negative) root that we keep as  $r_i$ . We have

$$\phi_i(t) = c_i e^{r_i t} + \phi_{i0} - \lambda_i^{-1} \tilde{g}_i \quad (76)$$

where  $\phi_{i0} - \lambda_i^{-1} \tilde{g}_i$  is the particular solution of Eq. 73, and  $c_i$  is a constant. We can determine  $c_i$  by enforcing the initial condition. Doing so, we find

$$\phi_i(t) = \phi_{i0} - \lambda_i^{-1} \tilde{g}_i (1 + e^{r_i t}). \quad (77)$$

Transforming back to  $\boldsymbol{\theta}_t$  space via the relationship  $\boldsymbol{\theta}_t = \mathbf{Q} \boldsymbol{\phi}_t$ , we have

$$\theta_i(t) = \sum_j Q_{ij} \phi_{j0} - Q_{ij} \lambda_j^{-1} \tilde{g}_j (1 + e^{r_j t}), \quad (78)$$



or equivalently

$$\boldsymbol{\theta}_t = \mathbf{Q}\phi_0 - \mathbf{Q}\boldsymbol{\Lambda}^{-1}(\mathbf{I} + e^{\mathbf{R}t})\tilde{\mathbf{g}} = \mathbf{Q}\phi_0 - \mathbf{Q}\boldsymbol{\Lambda}^{-1}\mathbf{Q}^T\mathbf{Q}(\mathbf{I} + e^{\mathbf{R}t})\mathbf{Q}^T\mathbf{Q}\tilde{\mathbf{g}}, \quad (79)$$

where we define the diagonal matrix  $\mathbf{R}$  whose diagonal entries are the  $r_i$ . Noting that

$$\mathbf{Q}(\mathbf{I} - e^{\mathbf{R}t})\mathbf{Q}^T = \mathbf{I} - \exp\left\{-\left[\sqrt{\frac{\gamma^2}{4} + \eta k \mathbf{H}} - \frac{\gamma}{2}\mathbf{I}\right]t\right\}, \quad (80)$$

we can simplify our result to

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 - \left(\mathbf{I} - \exp\left\{-\left[\sqrt{\frac{\gamma^2}{4} + \eta k \mathbf{H}} - \frac{\gamma}{2}\mathbf{I}\right]t\right\}\right)\mathbf{H}^{-1}\mathbf{g}. \quad (81)$$

This implies that, for any  $\Delta t > 0$ ,

$$\boldsymbol{\theta}_{\Delta t} - \boldsymbol{\theta}_0 = -\left(\mathbf{I} - \exp\left\{-\left[\sqrt{\frac{\gamma^2}{4} + \eta k \mathbf{H}} - \frac{\gamma}{2}\mathbf{I}\right]\Delta t\right\}\right)\mathbf{H}^{-1}\mathbf{g}. \quad (82)$$

## E.2 Special cases of the quadratic loss solution

If  $\Delta t$  is large, the exponential asymptotically vanishes, so

$$\lim_{\Delta t \rightarrow \infty} \boldsymbol{\theta}_{\Delta t} - \boldsymbol{\theta}_0 = -\mathbf{H}^{-1}\mathbf{g}, \quad (83)$$

and we recover Newton's method (i.e., one 'jumps' to the minimum). For large but not infinite  $\Delta t$ , Eq. 82 says that one ought to follow a Newton-like method, but with a possibly asymmetric learning rate in different directions, whose precise form depends on the argument of the exponential.

For small  $\Delta t$ , we obtain (to first order in  $\Delta t$ )

$$\boldsymbol{\theta}_{\Delta t} - \boldsymbol{\theta}_0 \approx \left[\sqrt{\frac{\gamma^2}{4} + \eta k \mathbf{H}} - \frac{\gamma}{2}\mathbf{I}\right]\mathbf{H}^{-1}\mathbf{g} \Delta t. \quad (84)$$

We can simplify this further if we make an assumption about the relative sizes of  $\mathbf{H}$  and  $\gamma$ . If  $\gamma \gg \sqrt{\eta k \lambda_i}$  for all eigenvalues  $\lambda_i$  of  $\mathbf{H}$ , then

$$\frac{\gamma}{2} \left[\sqrt{\mathbf{I} + \frac{4\eta k}{\gamma^2}\mathbf{H}} - \mathbf{I}\right] \approx \frac{\gamma}{2} \frac{2\eta k}{\gamma^2}\mathbf{H} = \frac{\eta k}{\gamma}\mathbf{H}, \quad (85)$$

so in this limit (the 'overdamped' limit) we obtain the learning rule

$$\boldsymbol{\theta}_{\Delta t} - \boldsymbol{\theta}_0 \approx \frac{\eta k}{\gamma}\mathbf{H}\mathbf{H}^{-1}\mathbf{g} \Delta t = \frac{\eta k}{\gamma}\mathbf{g} \Delta t, \quad (86)$$

which just corresponds to gradient descent.

Meanwhile, if  $\Delta t$  is small but  $\gamma \ll \sqrt{\eta k \lambda_i}$  for all eigenvalues  $\lambda_i$  of  $\mathbf{H}$  (or if  $\gamma = 0$ ), then

$$\boldsymbol{\theta}_{\Delta t} - \boldsymbol{\theta}_0 \approx \sqrt{\eta k \mathbf{H}}\mathbf{H}^{-1}\mathbf{g} \Delta t = \sqrt{\eta k}\mathbf{H}^{-1/2}\mathbf{g} \Delta t. \quad (87)$$

This 'ballistic' learning rule operates in a regime where momentum dominates learning dynamics.

Lastly, it's worth noting that if  $\gamma$  is larger than some eigenvalues of  $\mathbf{H}$  but smaller than others, one can obtain a learning rule that looks like gradient descent along some directions, and looks ballistic along other directions.



## F Deriving learning rules in non-Euclidean parameter spaces

In this appendix, we derive the results presented in the first part of Sec. 4. The relevant objective is

$$J(\{\boldsymbol{\theta}_t\}) = \int_0^\infty \left( \frac{1}{2\eta} \dot{\boldsymbol{\theta}}_t^T \mathbf{G}(\boldsymbol{\theta}_t) \dot{\boldsymbol{\theta}}_t + k\mathcal{L}(\boldsymbol{\theta}_t) \right) e^{-\gamma t} dt \quad (88)$$

where  $\mathbf{G}(\boldsymbol{\theta}_t)$  is a symmetric and positive definite matrix for all  $\boldsymbol{\theta}_t$ . The Lagrangian is

$$L(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}, t) = \left( \frac{1}{2\eta} \dot{\boldsymbol{\theta}}^T \mathbf{G}(\boldsymbol{\theta}) \dot{\boldsymbol{\theta}} + k\mathcal{L}(\boldsymbol{\theta}) \right) e^{-\gamma t} \quad (89)$$

and the EL equations are

$$\frac{d}{dt} \left( \mathbf{G}(\boldsymbol{\theta}_t) \dot{\boldsymbol{\theta}}_t \right) - \gamma \mathbf{G}(\boldsymbol{\theta}_t) \dot{\boldsymbol{\theta}}_t = \eta k \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t) + \nabla_{\boldsymbol{\theta}_t} \left( \frac{1}{2} \dot{\boldsymbol{\theta}}_t^T \mathbf{G}(\boldsymbol{\theta}_t) \dot{\boldsymbol{\theta}}_t \right). \quad (90)$$

If the metric is approximately independent of  $\boldsymbol{\theta}_t$ , we have the special form

$$\mathbf{G}(\boldsymbol{\theta}_t) \ddot{\boldsymbol{\theta}}_t - \gamma \mathbf{G}(\boldsymbol{\theta}_t) \dot{\boldsymbol{\theta}}_t \approx \eta k \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t) \implies \ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t \approx \eta k \mathbf{G}(\boldsymbol{\theta}_t)^{-1} \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t) \quad (91)$$

that appears in the main text. We can rewrite this in terms of the ‘momentum’  $\mathbf{p}_t := \dot{\boldsymbol{\theta}}_t$  to exactly reproduce the expression from Sec. 4.

### F.1 Solution to the Euler-Lagrange equations for a quadratic loss

Assume  $\mathbf{G}$  is independent of  $\boldsymbol{\theta}_t$  and that the loss is quadratic, i.e.,

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_t) := \mathcal{L}(\boldsymbol{\theta}_0) + \mathbf{g}^T (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^T \mathbf{H} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0), \quad (92)$$

where  $\mathbf{g}$  is the local gradient and  $\mathbf{H}$  is the (symmetric, positive definite) local Hessian. The EL equations become

$$\ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t = \eta k \mathbf{G}^{-1} [\mathbf{g} + \mathbf{H} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)]. \quad (93)$$

This is identical to what we considered in Appendix E, up to the changes  $\mathbf{g} \rightarrow \mathbf{G}^{-1} \mathbf{g}$  and  $\mathbf{H} \rightarrow \mathbf{G}^{-1} \mathbf{H}$ . The solution, then, is also identical up to these changes.



## G Non-gradient learning rules from partial controllability

In this appendix, we derive results related to the second part of Sec. 4, which concerns the influence of a ‘drift’ term (e.g., due to partial controllability) on optimal learning trajectories. The relevant objective is

$$J(\{\theta_t\}) = \int_0^\infty \left( \frac{1}{2\eta} [\dot{\theta}_t - \mathbf{f}(\theta_t)]^T \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] + k\mathcal{L}(\theta_t) \right) e^{-\gamma t} dt. \quad (94)$$

The Lagrangian is

$$L(\theta, \dot{\theta}, t) = \left( \frac{1}{2\eta} [\dot{\theta} - \mathbf{f}(\theta)]^T \mathbf{G}(\theta) [\dot{\theta} - \mathbf{f}(\theta)] + k\mathcal{L}(\theta) \right) e^{-\gamma t} \quad (95)$$

and the EL equations are

$$\frac{d}{dt} \left( \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] \right) - \gamma \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] = \eta k \nabla_{\theta_t} \mathcal{L}(\theta_t) + \nabla_{\theta_t} \left( \frac{1}{2} [\dot{\theta}_t - \mathbf{f}(\theta_t)]^T \mathbf{G}(\theta_t) [\dot{\theta}_t - \mathbf{f}(\theta_t)] \right).$$

**Trivial metric.** In the special case that  $\mathbf{G} = \mathbf{I}$ , we have

$$\ddot{\theta}_t - \frac{d}{dt} \mathbf{f}(\theta_t) - \gamma [\dot{\theta}_t - \mathbf{f}(\theta_t)] = \eta k \nabla_{\theta_t} \mathcal{L}(\theta_t) + \nabla_{\theta_t} \left( \frac{1}{2} \|\dot{\theta}_t - \mathbf{f}(\theta_t)\|^2 \right). \quad (96)$$

Recall that the Jacobian  $\mathbf{J}$  of  $\mathbf{f}$  is defined as<sup>8</sup>

$$J_{ij} := \frac{\partial f_i(\theta)}{\partial \theta_j}. \quad (97)$$

Using it, we can more explicitly write our expression as

$$\ddot{\theta}_t - \mathbf{J}(\theta_t) \dot{\theta}_t - \gamma [\dot{\theta}_t - \mathbf{f}(\theta_t)] = \eta k \nabla_{\theta_t} \mathcal{L}(\theta_t) - \mathbf{J}(\theta_t)^T [\dot{\theta}_t - \mathbf{f}(\theta_t)]. \quad (98)$$

Rearranging this, we obtain

$$\ddot{\theta}_t - [\gamma \mathbf{I} + \mathbf{J}(\theta_t) - \mathbf{J}(\theta_t)^T] \dot{\theta}_t = [\mathbf{J}(\theta_t)^T - \gamma \mathbf{I}] \mathbf{f}(\theta_t) + \eta k \nabla_{\theta_t} \mathcal{L}(\theta_t), \quad (99)$$

the equation that appears in the main text.

**Trivial metric and linear drift.** In the special case that  $\mathbf{f}(\theta) = \mathbf{J}\theta$ ,

$$\ddot{\theta}_t - [\gamma \mathbf{I} + \mathbf{J} - \mathbf{J}^T] \dot{\theta}_t = [\mathbf{J}^T - \gamma \mathbf{I}] \mathbf{J} \theta_t + \eta k \nabla_{\theta_t} \mathcal{L}(\theta_t). \quad (100)$$

In the case that  $\mathcal{L}$  is approximated as locally quadratic (as in, e.g., Appendix E), this becomes

$$\ddot{\theta}_t - [\gamma \mathbf{I} + \mathbf{J} - \mathbf{J}^T] \dot{\theta}_t = [\mathbf{J}^T - \gamma \mathbf{I}] \mathbf{J} \theta_t + \eta k [\mathbf{g} + \mathbf{H}(\theta_t - \theta_0)]. \quad (101)$$

In the next two subsections, we will consider two explicit examples of dynamics of this form.

### G.1 Example: drift term with rotational dynamics

**Setup.** For simplicity, assume that the Hessian is isotropic, i.e.,  $\mathbf{H} = h\mathbf{I}$ . Assume that the default dynamics are *rotational* in the sense that  $\mathbf{f}(\theta) = \mathbf{J}\theta$ , where  $\mathbf{J}$  is *skew-symmetric*, i.e.,  $\mathbf{J} = -\mathbf{J}^T$ . If  $\mathbf{J}$  is skew-symmetric, solutions of the default dynamics  $\dot{\theta} = \mathbf{f}(\theta)$  are purely rotational, since they have the form

$$\theta_t = e^{\mathbf{J}t} \theta_0 \quad (102)$$

where  $\exp(\mathbf{J}t)$  is a rotation matrix.

The EL equations have the form

$$\ddot{\theta}_t - [\gamma \mathbf{I} + 2\mathbf{J}] \dot{\theta}_t = -[\mathbf{J}^2 + \gamma \mathbf{J}] \theta_t + \eta k [\mathbf{g} + h(\theta_t - \theta_0)]. \quad (103)$$

---

<sup>8</sup>Note that we are using boldfaced  $\mathbf{J}$  to denote the Jacobian of  $\mathbf{f}$ , and  $J$  to denote the objective.



All skew-symmetric matrices are unitarily diagonalizable over the complex numbers, so we can write

$$\mathbf{J} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\dagger \quad (104)$$

where  $\mathbf{U}$  is unitary (i.e.,  $\mathbf{U} \mathbf{U}^\dagger = \mathbf{U}^\dagger \mathbf{U} = \mathbf{I}$ ) and  $\mathbf{\Lambda}$  is diagonal with complex entries.

Our second-order ODEs become decoupled in the space of eigenvectors of  $\mathbf{Q}$ . The quantity  $\phi_t := \mathbf{U}^\dagger \boldsymbol{\theta}_t$  changes according to

$$\ddot{\phi}_t - [\gamma \mathbf{I} + 2\mathbf{\Lambda}] \dot{\phi}_t = -[\mathbf{\Lambda}^2 + \gamma \mathbf{\Lambda}] \phi_t + \eta k [\tilde{\mathbf{g}} + h(\phi_t - \phi_0)] \quad (105)$$

where we define  $\tilde{\mathbf{g}} := \mathbf{U}^\dagger \mathbf{g}$ . Happily, each component  $\phi_i$  of  $\phi_t$  evolves independently according to a linear second-order ODE, and each of these can be solved in the usual way.

**Decoupled ODEs.** Each  $\phi_i$  evolves according to

$$\ddot{\phi}_i - [\gamma + 2\lambda_i] \dot{\phi}_i - [\eta k h - \lambda_i^2 - \gamma \lambda_i] \phi_i = \eta k [\tilde{g}_i - h \phi_{i0}] \quad (106)$$

where  $\lambda_i := \Lambda_{ii}$ . The characteristic equation of this ODE is

$$r^2 - [\gamma + 2\lambda_i]r - [\eta k h - \lambda_i^2 - \gamma \lambda_i] = 0 \quad (107)$$

and its solution is

$$\begin{aligned} r_{\pm} &= \frac{\gamma + 2\lambda_i}{2} \pm \sqrt{\frac{(\gamma + 2\lambda_i)^2}{4} + \eta k h - \lambda_i^2 - \gamma \lambda_i} \\ &= \lambda_i + \frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} + \eta k h} . \end{aligned} \quad (108)$$

As before (see, e.g., Appendix C and Appendix E), we ignore the positive root since its associated solution asymptotically blows up. Define

$$r_i := \lambda_i + \frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} + \eta k h} \quad (109)$$

and  $\tilde{\mathbf{R}}$  as the diagonal matrix with  $\tilde{R}_{ii} = r_i$ . Combining the general and particular solutions of Eq. 106 yields

$$\phi_i(t) = \frac{\eta k h}{\eta k h - \lambda_i(\lambda_i + \gamma)} [\phi_{i0} - h^{-1} \tilde{g}_i] + c e^{r_i t} \quad (110)$$

for some constant  $c$ . Enforcing the initial condition,

$$\phi_i(t) = \frac{\eta k h}{\eta k h - \lambda_i(\lambda_i + \gamma)} [\phi_{i0} - h^{-1} \tilde{g}_i] (1 - e^{r_i t}) + \phi_{i0} e^{r_i t} . \quad (111)$$

Define the diagonal matrix  $\tilde{\mathbf{B}}$  via

$$\tilde{B}_{ii} := \frac{\eta k h}{\eta k h - \lambda_i(\lambda_i + \gamma)} , \quad (112)$$

so that the  $\phi_i$  solution becomes

$$\phi_i(t) = (1 - e^{r_i t}) \tilde{B}_i (\phi_{i0} - h^{-1} \tilde{g}_i) + e^{r_i t} \phi_{i0} , \quad (113)$$

or in vector form,

$$\phi_t = (\mathbf{I} - e^{\tilde{\mathbf{R}} t}) \tilde{\mathbf{B}} (\phi_0 - h^{-1} \tilde{\mathbf{g}}) + e^{\tilde{\mathbf{R}} t} \phi_0 . \quad (114)$$

**Solution.** The relationship  $\boldsymbol{\theta}_t = \mathbf{U} \phi_t$  tells us that

$$\boldsymbol{\theta}_t = e^{\mathbf{R} t} \boldsymbol{\theta}_0 + (\mathbf{I} - e^{\mathbf{R} t}) \mathbf{B} (\boldsymbol{\theta}_0 - h^{-1} \mathbf{g}) \quad (115)$$

where we define the matrix  $\mathbf{R}$  as

$$\mathbf{R} := \mathbf{U} \tilde{\mathbf{R}} \mathbf{U}^\dagger = \mathbf{J} + \left( \frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} + \eta k h} \right) \mathbf{I} \quad (116)$$



and  $B$  as

$$B := U \tilde{B} U^\dagger = \eta k h [\eta k h - \mathbf{J}(\mathbf{J} + \gamma \mathbf{I})]^{-1}. \quad (117)$$

Interestingly, in the long-time limit this solution doesn't converge to the global minimum, but to a slightly different location controlled by the 'bias' matrix  $B$ :

$$\lim_{t \rightarrow \infty} \boldsymbol{\theta}_t = B(\boldsymbol{\theta}_0 - h^{-1} \mathbf{g}). \quad (118)$$

An even more interesting feature of this solution is that, since the  $r_i$  are complex-valued (through their dependence on the  $\lambda_i$ , which are pure imaginary for skew-symmetric matrices), they approach their asymptotic value in a spiraling-in fashion. To see this explicitly, consider the particular skew-symmetric matrix

$$\mathbf{J} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad (119)$$

which is the infinitesimal generator of a counterclockwise rotation in a two-dimensional plane. Then

$$e^{\mathbf{J}t} = \cos(t) \mathbf{I} + \sin(t) \mathbf{J} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}, \quad (120)$$

i.e., we obtain a matrix that performs a counterclockwise rotation by an angle  $t$ . The optimal learning trajectory approaches its final value according to

$$e^{\mathbf{R}t} = e^{\mathbf{J}t} e^{-\left(\sqrt{\frac{\gamma^2}{4} + \eta k h - \frac{\gamma}{2}}\right)t}, \quad (121)$$

which combines decay with rotation.

## G.2 Example: drift term representing weight decay

**Setup.** In this example, assume a general Hessian  $\mathbf{H}$  and that  $\mathbf{f}(\boldsymbol{\theta}) = -j\boldsymbol{\theta}$  for some weight decay rate  $j > 0$ . The EL equations are

$$\ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t = (j - \gamma)j \boldsymbol{\theta}_t + \eta k [\mathbf{g} + \mathbf{H}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)], \quad (122)$$

or equivalently

$$\ddot{\boldsymbol{\theta}}_t - \gamma \dot{\boldsymbol{\theta}}_t - [(j - \gamma)j \mathbf{I} + \eta k \mathbf{H}] \boldsymbol{\theta}_t = \eta k (\mathbf{g} - \mathbf{H} \boldsymbol{\theta}_0). \quad (123)$$

We can solve this system of second-order ODEs using the strategy from Appendix E, by using an eigendecomposition  $\mathbf{H} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T$ . The only differences are that the relevant rates are

$$r_i := \frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} + (j - \gamma)j + \eta k \lambda_i} \quad (124)$$

and that we once again get biases

$$b_i := \frac{\eta k \lambda_i}{\eta k \lambda_i + (j - \gamma)j}. \quad (125)$$

Define the diagonal matrix  $\tilde{\mathbf{R}}$  using the  $r_i$ , and the diagonal matrix  $\tilde{\mathbf{B}}$  using the  $b_i$ . The solution has

$$\boldsymbol{\theta}_t = e^{\mathbf{R}t} \boldsymbol{\theta}_0 + (\mathbf{I} - e^{\mathbf{R}t}) \mathbf{B}(\boldsymbol{\theta}_0 - \mathbf{H}^{-1} \mathbf{g}) \quad (126)$$

where we define the matrix  $\mathbf{R}$  as

$$\mathbf{R} := \mathbf{Q} \tilde{\mathbf{R}} \mathbf{Q}^T = \frac{\gamma}{2} \mathbf{I} - \sqrt{\left(\frac{\gamma^2}{4} + (j - \gamma)j\right) \mathbf{I} + \eta k \mathbf{H}} \quad (127)$$

and  $\mathbf{B}$  as

$$\mathbf{B} := \mathbf{Q} \tilde{\mathbf{B}} \mathbf{Q}^T = \frac{\eta k \mathbf{H}}{\eta k \mathbf{H} + (j - \gamma)j}. \quad (128)$$

As in the previous example, the form of  $\mathbf{J}$  (here, just the scalar  $j$ ) contributes to both a drift-related bias  $\mathbf{B}$  and the rate  $\mathbf{R}$  at which the optimal trajectory approaches its asymptotic value.



### G.3 Optimal learning trajectories generally do not follow gradients

The  $\mathbf{G} = \mathbf{I}$  EL equations above can be written as

$$\dot{\mathbf{p}}_t = [\gamma \mathbf{I} + \mathbf{J}(\boldsymbol{\theta}_t) - \mathbf{J}(\boldsymbol{\theta}_t)^T] \mathbf{p}_t + [\mathbf{J}(\boldsymbol{\theta}_t)^T - \gamma \mathbf{I}] \mathbf{f}(\boldsymbol{\theta}_t) + \eta k \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t), \quad (129)$$

where we define momentum (as before) as  $\mathbf{p}_t := \dot{\boldsymbol{\theta}}_t$ . Assume Helmholtz decomposition  $\mathbf{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) + \mathbf{R}(\boldsymbol{\theta})$ , where  $V$  is some non-unique ‘potential’ function and  $\mathbf{R}$  is divergence-free (i.e.,  $\nabla_{\boldsymbol{\theta}} \cdot \mathbf{R} = 0$ ). This decomposition implies that the Jacobian of  $\mathbf{f}$  has entries

$$J_{ij} = \partial_{ij}^2 V(\boldsymbol{\theta}) + \frac{\partial R_i(\boldsymbol{\theta})}{\partial \theta_j}. \quad (130)$$

Note that the only possible source of asymmetry comes from  $\mathbf{R}$ . Moreover, note that

$$\begin{aligned} \sum_j J_{ij}^T f_j &= \sum_j \left[ \partial_{ij}^2 V + \frac{\partial R_j}{\partial \theta_i} \right] [\partial_j V + R_j] = \partial_i \left( \sum_j \frac{(\partial_j V)^2}{2} + (\partial_j V) R_j + \frac{R_j^2}{2} \right) \\ &= \partial_i \left( \sum_j \frac{[\partial_j V + R_j]^2}{2} \right). \end{aligned} \quad (131)$$

If  $\mathbf{J}_{\mathbf{R}}$  denotes the Jacobian of  $\mathbf{R}$ , the EL equations can be written

$$\begin{aligned} \dot{\mathbf{p}}_t &= [\gamma \mathbf{I} + \mathbf{J}_{\mathbf{R}}(\boldsymbol{\theta}_t) - \mathbf{J}_{\mathbf{R}}(\boldsymbol{\theta}_t)^T] \mathbf{p}_t - \gamma \nabla_{\boldsymbol{\theta}_t} V(\boldsymbol{\theta}_t) - \gamma \mathbf{R}(\boldsymbol{\theta}_t) + \mathbf{J}(\boldsymbol{\theta}_t)^T \mathbf{f}(\boldsymbol{\theta}_t) + \eta k \nabla_{\boldsymbol{\theta}_t} \mathcal{L}(\boldsymbol{\theta}_t) \\ &= [\gamma \mathbf{I} + \mathbf{J}_{\mathbf{R}}(\boldsymbol{\theta}_t) - \mathbf{J}_{\mathbf{R}}(\boldsymbol{\theta}_t)^T] \mathbf{p}_t - \gamma \mathbf{R}(\boldsymbol{\theta}_t) + \nabla_{\boldsymbol{\theta}_t} V_{eff}(\boldsymbol{\theta}_t) \end{aligned}$$

where we define the *effective loss/potential*

$$V_{eff}(\boldsymbol{\theta}_t) := \eta k \mathcal{L}(\boldsymbol{\theta}_t) + \frac{1}{2} \|\nabla_{\boldsymbol{\theta}_t} V(\boldsymbol{\theta}_t) + \mathbf{R}\|^2 - \gamma V(\boldsymbol{\theta}_t). \quad (132)$$

Hence, it is clear that a nontrivial  $\mathbf{R}$  contributes non-gradient dynamics to learning in two ways: first, through determining an asymmetric effective temporal discounting rate; and second, through the  $\gamma \mathbf{R}$  term.



## H Deriving adaptive learning rules from dynamic loss landscape beliefs

In this appendix, we motivate and derive the Adam-like adaptive learning rule discussed in Sec. 5.

### H.1 Motivation: why gradient variance relates to the Hessian

First, we briefly motivate that the (average) loss landscape can be locally approximated as

$$\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_t)] = \mathcal{L}(\boldsymbol{\theta}_0) + \mathbf{m}_t^T(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) + \frac{\kappa}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^T \mathbf{V}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) \quad (133)$$

where  $\mathbf{m}_t$  is the average observed gradient and  $\mathbf{V}_t$  is the covariance of these gradient observations. Why is it reasonable to suppose that  $\mathbf{V}_t$  is proportional to the local Hessian  $\mathbf{H}$ ?

Assume that the true loss landscape has a gradient  $\mathbf{g}$  and Hessian  $\mathbf{H}$ . The idea is to view noise in gradients as due to an unobservable, noisy parameter vector  $\tilde{\boldsymbol{\theta}}_t$  that explores the local loss landscape according to a stochastic process. Since  $\tilde{\boldsymbol{\theta}}_t$  is noisy, its time derivative (i.e., gradients, which are observable) will also be noisy. We can view this as a change in perspective. Rather than assuming that  $\boldsymbol{\theta}_t$  remains fixed but the landscape changes dynamically (and partly randomly) around us, we can assume that the landscape is fixed, but that *where we are in it* is randomly changing.

The simplest process we can assume is one that is unbiased (i.e., noisy gradients equal true gradients on average), and has a structureless (i.e., white and state-independent) noise term. Consider

$$\frac{d}{dt}\tilde{\boldsymbol{\theta}}_t = -\frac{1}{\tau}\nabla_{\tilde{\boldsymbol{\theta}}_t}\mathcal{L} + \sigma\boldsymbol{\eta}_t = -\frac{1}{\tau}\left[\mathbf{g} + \mathbf{H}(\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)\right] + \sigma\boldsymbol{\eta}_t = \frac{1}{\tau}\mathbf{H}\left[\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}}_t\right] + \sigma\boldsymbol{\eta}_t \quad (134)$$

where  $\tau > 0$  is a decay time scale,  $\sigma > 0$  controls the amount of noise,  $\boldsymbol{\eta}_t$  is a Gaussian white noise term, and

$$\boldsymbol{\mu} := \boldsymbol{\theta}_0 - \mathbf{H}^{-1}\mathbf{g}. \quad (135)$$

At steady state (or at least, on time scales somewhat longer than  $\tau$ ), we have

$$\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{2}\mathbf{H}^{-1}). \quad (136)$$

Near steady state, this implies that

$$\mathbb{E}\{(\nabla_{\tilde{\boldsymbol{\theta}}_t}\mathcal{L})(\nabla_{\tilde{\boldsymbol{\theta}}_t}\mathcal{L})^T\} = \frac{1}{\tau^2}\mathbf{H}\mathbb{E}\{(\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}}_t)(\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}}_t)^T\}\mathbf{H} = \frac{\sigma^2}{2\tau^2}\mathbf{H} \quad (137)$$

or equivalently that the variance of observed gradients is proportional to  $\mathbf{H}$ . This result can be interpreted in the following way. For sharp/curved minima,  $\mathbf{H}$  is by definition large, so small parameter changes can produce relatively large changes in gradients; on the other hand, in flat loss landscape regions,  $\mathbf{H}$  is small, so even large parameter changes do not change gradients much.

### H.2 Deriving the Euler-Lagrange equations

**More complex observation model.** Assume that  $\mathbf{g}_t \sim \mathcal{N}(\mathbf{m}_t, \mathbf{V}_t/\Delta t)$ . The relevant objective is

$$J(\{\boldsymbol{\theta}_t, \mathbf{m}_t, \mathbf{v}_t\}) = \lim_{\Delta t \rightarrow 0} \int_0^\infty \left( \frac{\|\boldsymbol{\theta}_t\|^2}{2\eta} - \frac{\log p(\mathbf{m}_{t+1}, \mathbf{V}_{t+1}|\mathbf{g}_t, \mathbf{m}_t, \mathbf{V}_t)}{\Delta t} + k\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_t)] \right) e^{-\gamma t} dt$$

where  $p(\mathbf{m}_{t+1}, \mathbf{V}_{t+1}|\mathbf{g}_t, \mathbf{m}_t, \mathbf{V}_t)$  is the learner's posterior belief about local landscape shape dynamics. The posterior term contains two types of terms: terms from the observation model, and terms from the prior. In particular, it contains the terms

$$\sum_i \frac{(g_i - m_i)^2}{2v_i} + \frac{1}{2} \log(2\pi v_i) + \frac{\|\dot{\mathbf{m}}_t + \alpha_1 \mathbf{m}_t\|^2}{2\xi_1^2} + \frac{\|\dot{\mathbf{v}}_t + \alpha_2 \mathbf{v}_t\|^2}{2\xi_2^2} + \text{const.} \quad (138)$$

where we neglect unimportant additive constants, and assume that  $\mathbf{V}_t$  is diagonal. Here,  $v_i$  denotes  $V_{ii}$ , and  $\mathbf{v}$  denotes the vector containing the  $v_i$ . The corresponding Lagrangian is

$$L := \sum_i \left( \frac{\dot{\theta}_i^2}{2\eta} + \frac{(g_i - m_i)^2}{2v_i} + \frac{1}{2} \log(2\pi v_i) + \frac{(\dot{m}_i + \alpha_1 m_i)^2}{2\xi_1^2} + \frac{(\dot{v}_i + \alpha_2 v_i)^2}{2\xi_2^2} + k\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_t)] \right) e^{-\gamma t},$$



and the corresponding EL equations are

$$\begin{aligned}
\ddot{\theta}_i - \gamma \dot{\theta}_i &= \eta k [g_i + \kappa v_i (\theta_i - \theta_{0i})] \\
\ddot{m}_i - \gamma (\dot{m}_i + \alpha_1 m_i) &= \alpha_1^2 m_i + \xi_1^2 \left[ k(\theta_i - \theta_{0i}) + \frac{(m_i - g_i)}{v_i} \right] \\
\ddot{v}_i - \gamma (\dot{v}_i + \alpha_2 v_i) &= \alpha_2^2 v_i + \xi_2^2 \left[ k \frac{\kappa}{2} (\theta_i - \theta_{0i})^2 + \frac{1}{2} \frac{1}{v_i} - \frac{1}{2} \frac{(g_i - m_i)^2}{2v_i^2} \right].
\end{aligned} \tag{139}$$

The equation for  $v_i$  is somewhat more complicated than the one presented in the main text, but note that it can be written as

$$\ddot{v}_i - \gamma (\dot{v}_i + \alpha_2 v_i) = \alpha_2^2 v_i + \xi_2^2 \left[ k \frac{\kappa}{2} (\theta_i - \theta_{0i})^2 + \frac{1}{2} \frac{1}{v_i^2} (v_i - (g_i - m_i)^2) \right], \tag{140}$$

which matches the main text form up to the prefactor  $1/(2v_i^2)$ . If the variance  $v_i$  is fairly stable (for example, because a reasonable amount of evidence about gradients has already been accumulated), then this prefactor is approximately constant, and the forms are identical.

**Simplified observation model.** If we instead treat gradients and squares of gradients as consisting of separate observations, i.e.,  $\mathbf{g}_t \sim \mathcal{N}(\mathbf{m}_t, (\sigma_1^2/\Delta t)\mathbf{I})$  and  $(\mathbf{g}_t - \mathbf{m}_t)(\mathbf{g}_t - \mathbf{m}_t)^T \sim \mathcal{N}(\mathbf{V}_t, (\sigma_2^2/\Delta t)\mathbf{I})$ , then the posterior contains the terms

$$\sum_i \frac{(g_i - m_i)^2}{2\sigma_1^2} + \frac{[v_i - (g_i - m_i)^2]^2}{2\sigma_2^2} + \frac{\|\dot{\mathbf{m}}_t + \alpha_1 \mathbf{m}_t\|^2}{2\xi_1^2} + \frac{\|\dot{\mathbf{v}}_t + \alpha_2 \mathbf{v}_t\|^2}{2\xi_2^2} + \text{const.} \tag{141}$$

The corresponding Lagrangian is

$$L := \sum_i \left( \frac{\dot{\theta}_i^2}{2\eta} + \frac{(g_i - m_i)^2}{2\sigma_1^2} + \frac{[v_i - (g_i - m_i)^2]^2}{2\sigma_2^2} + \frac{(\dot{m}_i + \alpha_1 m_i)^2}{2\xi_1^2} + \frac{(\dot{v}_i + \alpha_2 v_i)^2}{2\xi_2^2} + k\mathbb{E}[\hat{\mathcal{L}}(\boldsymbol{\theta}_t)] \right) e^{-\gamma t},$$

and the corresponding EL equations are

$$\begin{aligned}
\ddot{\theta}_i - \gamma \dot{\theta}_i &= \eta k [g_i + \kappa v_i (\theta_i - \theta_{0i})] \\
\ddot{m}_i - \gamma (\dot{m}_i + \alpha_1 m_i) &= \alpha_1^2 m_i + \xi_1^2 \left[ k(\theta_i - \theta_{0i}) + \frac{(m_i - g_i)}{\sigma_1^2} \right] \\
\ddot{v}_i - \gamma (\dot{v}_i + \alpha_2 v_i) &= \alpha_2^2 v_i + \xi_2^2 \left[ k \frac{\kappa}{2} (\theta_i - \theta_{0i})^2 + \frac{1}{\sigma_2^2} (v_i - (g_i - m_i)^2) \right].
\end{aligned} \tag{142}$$



## I Deriving learning rules sensitive to weight uncertainty

In this appendix, we derive the weight-uncertainty-sensitive learning rule discussed in Sec. 6. Recall that we assume each model parameter  $\theta_i$  (for  $i = 1, \dots, D$ ) is associated with a normal distribution  $\mathcal{N}(\mu_i, v_i)$ , where  $\mu_i$  is the average value of  $\theta_i$ , and  $v_i$  is its variance. The advantage of this setup is that it allows the learner to not just estimate what their parameters are, but also how certain they are about them. In principle, we could consider a model with a more general covariance matrix, but we restrict ourselves to a diagonal covariance for simplicity.

### I.1 Simplifying the objective

The relevant objective is

$$J(\{\boldsymbol{\mu}_t, \mathbf{v}_t\}) = \lim_{\Delta t \rightarrow 0} \int_0^\infty \left[ \frac{D_{KL}(p(\boldsymbol{\theta}|\boldsymbol{\mu}_{t+1}, \mathbf{v}_{t+1})||p(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \mathbf{v}_t))}{\eta(\Delta t)^2} - \mathcal{H}(p(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \mathbf{v}_t)) + k\mathcal{L}(\boldsymbol{\mu}_t, \mathbf{v}_t) \right] e^{-\gamma t} dt$$

where the first term does not penalize abrupt parameter changes, but abrupt changes in parameter distribution. Note that the Kullback-Leibler (KL) divergence term can be written

$$\begin{aligned} & D_{KL}(p(\boldsymbol{\theta}|\boldsymbol{\mu}_t + \dot{\boldsymbol{\mu}}_t \Delta t, \mathbf{v}_t + \dot{\mathbf{v}}_t \Delta t)||p(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \mathbf{v}_t)) \\ &= \mathbb{E}_{\boldsymbol{\theta}} \{ \log p(\boldsymbol{\theta}|\boldsymbol{\mu}_t + \dot{\boldsymbol{\mu}}_t \Delta t, \mathbf{v}_t + \dot{\mathbf{v}}_t \Delta t) - \log p(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \mathbf{v}_t) \} \\ &= \sum_i \mathbb{E}_{\theta_i} \left\{ -\frac{[\theta_i - \mu_i - \dot{\mu}_i \Delta t]^2}{2(v_i + \dot{v}_i \Delta t)} - \frac{1}{2} \log[2\pi(v_i + \dot{v}_i \Delta t)] + \frac{[\theta_i - \mu_i]^2}{2v_i} + \frac{1}{2} \log[2\pi v_i] \right\} \\ &= \sum_i -\frac{1}{2} - \frac{1}{2} \log[2\pi(v_i + \dot{v}_i \Delta t)] + \frac{v_i + \dot{v}_i \Delta t}{2v_i} + \frac{\dot{\mu}_i^2}{2v_i} (\Delta t)^2 + \frac{1}{2} \log[2\pi v_i] \\ &= \sum_i \frac{\dot{v}_i \Delta t}{2v_i} + \frac{\dot{\mu}_i^2}{2v_i} (\Delta t)^2 - \frac{1}{2} \log \left( 1 + \frac{\dot{v}_i}{v_i} \Delta t \right) \\ &\approx \sum_i \left( \frac{1}{2} \frac{\dot{\mu}_i^2}{v_i} + \frac{1}{4} \frac{\dot{v}_i^2}{v_i^2} \right) (\Delta t)^2 \end{aligned}$$

and that the entropy term is

$$\mathcal{H}(p(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \mathbf{v}_t)) = \sum_i \frac{1}{2} \log(2\pi e v_i), \quad (143)$$

which means that our objective is effectively

$$J(\{\boldsymbol{\mu}_t, \mathbf{v}_t\}) = \int_0^\infty \left[ \sum_i \frac{1}{2\eta} \frac{\dot{\mu}_i^2}{v_i} + \frac{1}{4\eta} \frac{\dot{v}_i^2}{v_i^2} - \frac{1}{2} \log(2\pi e v_i) + k\mathcal{L}(\boldsymbol{\mu}_t, \mathbf{v}_t) \right] e^{-\gamma t} dt$$

and the corresponding Lagrangian is

$$L(\boldsymbol{\mu}, \mathbf{v}, \dot{\boldsymbol{\mu}}, \dot{\mathbf{v}}, t) = \left[ \sum_i \frac{1}{2\eta} \frac{\dot{\mu}_i^2}{v_i} + \frac{1}{4\eta} \frac{\dot{v}_i^2}{v_i^2} - \frac{1}{2} \log(2\pi e v_i) + k\mathcal{L}(\boldsymbol{\mu}_t, \mathbf{v}_t) \right] e^{-\gamma t}.$$

Here, the effective metric is the Fisher information metric for a normal distribution.

### I.2 Simplifying the Euler-Lagrange equations

Taking derivatives, the EL equations read

$$\begin{aligned} & \frac{d}{dt} \left( \frac{\dot{\mu}_i}{v_i} \right) - \gamma \frac{\dot{\mu}_i}{v_i} = \eta k \frac{\partial \mathcal{L}(\boldsymbol{\mu}, \mathbf{v})}{\partial \mu_i} \\ & \frac{d}{dt} \left( \frac{1}{2} \frac{\dot{v}_i}{v_i^2} \right) - \gamma \frac{1}{2} \frac{\dot{v}_i}{v_i^2} = \eta k \frac{\partial \mathcal{L}(\boldsymbol{\mu}, \mathbf{v})}{\partial v_i} - \frac{\eta}{2} \frac{1}{v_i} - \frac{\dot{\mu}_i^2}{2v_i^2} - \frac{\dot{v}_i^2}{2v_i^3}. \end{aligned} \quad (144)$$



Simplifying, these become

$$\begin{aligned}\ddot{\mu}_i - \left[ \frac{\dot{v}_i}{v_i} + \gamma \right] \dot{\mu}_i &= \eta k v_i \frac{\partial \mathcal{L}(\boldsymbol{\mu}, \mathbf{v})}{\partial \mu_i} \\ \ddot{v}_i - \gamma \dot{v}_i &= 2\eta \left[ k v_i^2 \frac{\partial \mathcal{L}(\boldsymbol{\mu}, \mathbf{v})}{\partial v_i} - \frac{1}{2} v_i \right] + \frac{\dot{v}_i^2}{v_i} - \dot{\mu}_i^2.\end{aligned}\tag{145}$$

Consider a local (quadratic) approximation of the loss, i.e.,

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_t) := \mathcal{L}(\boldsymbol{\theta}_0) + \mathbf{g}^T(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^T \mathbf{H}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0),\tag{146}$$

where  $\mathbf{g}$  is the local gradient and  $\mathbf{H}$  is the local Hessian. Averaging this quantity over  $\boldsymbol{\theta}_t$  and  $\boldsymbol{\theta}_0$ ,

$$\mathcal{L}(\boldsymbol{\mu}, \mathbf{v}) := \mathbb{E}_{\boldsymbol{\theta}_t, \boldsymbol{\theta}_0} \{\hat{\mathcal{L}}(\boldsymbol{\theta}_t)\} := \mathbf{g}^T(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0) + \frac{1}{2}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^T \mathbf{H}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0) + \sum_i \frac{1}{2} H_{ii} v_i + \text{const.}$$

where ‘const.’ denotes terms we can ignore. Using this particular  $\mathcal{L}(\boldsymbol{\mu}, \mathbf{v})$ , the EL equations become

$$\begin{aligned}\ddot{\mu}_i - \left[ \frac{\dot{v}_i}{v_i} + \gamma \right] \dot{\mu}_i &= \eta k v_i \left[ g_i + \sum_j H_{ij}(\mu_j - \mu_{j0}) \right] \\ \ddot{v}_i - \gamma \dot{v}_i &= 2\eta \left[ \frac{k}{2} H_{ii} v_i^2 - \frac{1}{2} v_i \right] + \frac{\dot{v}_i^2}{v_i} - \dot{\mu}_i^2,\end{aligned}\tag{147}$$

which are the equations that appear in the main text.

### I.3 The Euler-Lagrange equations in the overdamped limit

The EL equations simplify somewhat in the overdamped (large  $\gamma$ ) limit<sup>9</sup>:

$$\begin{aligned}\dot{\mu}_i &\approx -\frac{\eta k v_i}{\gamma} \left[ g_i + \sum_j H_{ij}(\mu_j - \mu_{j0}) \right] \\ \dot{v}_i &\approx \frac{\eta}{\gamma} [v_i - k H_{ii} v_i^2].\end{aligned}\tag{148}$$

Note that, in this limit,  $v_i$  influences how  $\mu_i$  evolves (by modulating the effective learning rate), but  $v_i$  evolves independently of  $\mu_i$ , at least on time scales where the quadratic loss landscape approximation remains valid. The entropic term  $v_i$  tends to increase the variance, and the loss gradient term tends to decrease the variance; they balance when

$$v_i = \frac{1}{k H_{ii}},\tag{149}$$

which implies that narrower loss landscape basins (high  $H_{ii}$ ) produce small uncertainties, and broader basins (low  $H_{ii}$ ) produce high uncertainties. This is intuitively reasonable, and reflects the selection of the ‘simplest’ model compatible with the data. Meanwhile, when the entropic term is absent, in this limit  $v_i \rightarrow 0$ .

Suppose that the learner has converged to a ‘good’ (i.e., deep) local or global minimum. When there is a task transition, one generally expects the local landscape to no longer be as curved (since the learner is probably no longer in a good local minimum), which means that  $H_{ii}$  suddenly decreases. Eq. 148 says that optimal learning dynamics involves a sudden increase in variance, which persists until  $v_i$  equilibrates to its new value. By Eq. 149, this value is proportional to the new  $1/H_{ii}$ .

<sup>9</sup>To justify this reduction more rigorously, we could have performed a singular perturbation analysis. This analysis is not particularly enlightening, so we merely report the result.



#### I.4 Qualitative behavior outside of the overdamped limit

Outside of the overdamped limit, the EL equations for  $\mu_i$  and  $v_i$  influence each other: changes in  $\mu_i$  contribute an effective ‘force’ that affects  $v_i$ , and  $v_i$  affects both the effective discounting rate and learning rate in the  $\mu_i$  equation. Because the  $\dot{\mu}_i^2$  term has the same sign as the entropic term, it plays the same qualitative role, and can produce increases in variance.

Again suppose that there is a task transition after the learner has converged to a ‘good’ minimum, so that  $H_{ii}$  suddenly decreases. Optimal learning dynamics again involves a sudden increase in variance, this time driven by both the entropic term and the  $\dot{\mu}_i^2$  term.

Unlike in the overdamped case, if  $H_{ii}$  does not change after a task transition but  $g_i$  *does* (i.e., the location, but not size, of the basin has changed), the  $\dot{\mu}_i^2$  produces a transient increase in variance. After  $\mu_i$  equilibrates, the  $\dot{\mu}_i^2$  term becomes zero, and  $v_i$  approaches the same equilibrium value as in the overdamped case (Eq. 149).