Adaptive Learning Methods

Karl Stratos

Last updated: March, 2026

Contents

1 Online Convex Optimization (OCO)

At step $t = 1, 2, \ldots$, we propose $w_t \in V \subseteq \mathbb{R}^d$ where V is closed and convex. The enemy then chooses a convex and differentiable loss $l_t : \mathbb{R}^d \to \mathbb{R}$ and punishes us by $l_t(w_t) \in \mathbb{R}$. Assuming T such steps, let $u = \arg\min_{w \in V} \sum_{t=1}^{T} l_t(w)$ denote the best hypothesis in retrospect. We want to upper bound the total "regret" as a function of T:

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq B(T)$$

The goal is to develop an algorithm for achieving a sublinear regret bound $B(T) = o(T)$. A sublinear OCO algorithm can be used to solve CSO (see Appendix A). A well-known lower bound is $\Omega(\sqrt{T})$; there exist environments where no algorithm can achieve regret smaller than $\sqrt{T}$ asymptotically (see Appendix B).

1.1 Mirror Descent

Let $\hat{l}_t(w) = l_t(w_t) + g_t^\top (w - w_t)$ denote the linearization of the loss around w_t where $g_t = \nabla l_t(w_t)$. To make the minimum finite, we regularize by the Bregman divergence $D_{\psi_t}(\cdot, w_t)$ (Appendix C) where $\psi_t : V \to \mathbb{R}$ is strictly convex and differentiable. Assuming $\eta_t > 0$, our per-step objective is

$$w_{t+1} = \arg\min_{w \in V} \ \hat{l}_t(w) + \frac{1}{\eta_t} D_{\psi_t}(w, w_t) \tag{1}$$

For instance, (1) becomes gradient descent when $V = \mathbb{R}^d$ and $\psi_t(w) = \frac{1}{2} ||w||_2^2$ and exponentiated gradient descent when $V = \Delta^{d-1}$ and $\psi_t(w) = -H(w)$ (Appendix D). Instead of computing (1) directly, we may perform unconstrained minimization and project:

$$\tilde{w}_{t+1} = \arg\min_{w \in \mathbb{R}^d} \ \eta_t \hat{l}_t(w) + D_{\psi_t}(w, w_t) = \nabla \psi_t^*(\nabla \psi_t(w_t) - \eta_t g_t) \tag{2}$$

$$w_{t+1} = \arg\min_{w \in V} \ D_{\psi_t}(w, \tilde{w}_{t+1}) \tag{3}$$

where $\psi_t^* : \mathbb{R}^d \to \mathbb{R}$ is the convex conjugate of ψ_t (Fact E.4). We can show that (1) and (3) are equal.[1] We call this approach **online mirror descent** because the current hypothesis is "mirrored" to a dual coordinate system $z_t = \nabla \psi_t(w_t)$, updated by standard gradient descent $z_{t+1} = z_t - \eta_t g_t$, and mirrored back $\tilde{w}_{t+1} = \nabla \psi_t^*(z_{t+1})$ to be projected to V.

1.2 General Analysis

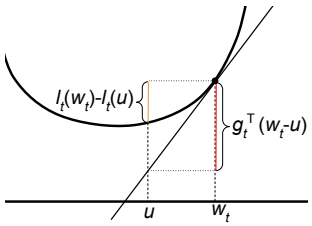
Since (1) is the Bregman projection of w_t onto V regularized by $\eta_t \hat{l}_t$, the Pythagorean theorem gives us $D_{\psi_t}(u, w_t) + \eta_t \hat{l}_t(u) \geq D_{\psi_t}(w_{t+1}, w_t) + D_{\psi_t}(u, w_{t+1}) + \eta_t \hat{l}_t(w_{t+1})$ (Lemma C.3), or

$$\begin{aligned}
D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1}) &\geq \eta_t \hat{l}_t(w_{t+1}) - \eta_t \hat{l}_t(u) + D_{\psi_t}(w_{t+1}, w_t) \\
&\geq \eta_t g_t^\top (w_t - u) + \eta_t g_t^\top (w_{t+1} - w_t) + \frac{\sigma_t}{2} ||w_{t+1} - w_t||_t^2 && (4) \\
&\geq \eta_t g_t^\top (w_t - u) - \eta_t ||g_t||_{t,*} ||w_{t+1} - w_t||_t + \frac{\sigma_t}{2} ||w_{t+1} - w_t||_t^2 && (5) \\
&\geq \eta_t g_t^\top (w_t - u) - \frac{\eta_t^2}{2\sigma_t} ||g_t||_{t,*}^2 && (6) \\
&\geq \eta_t \left(l_t(w_t) - l_t(u)\right) - \frac{\eta_t^2}{2\sigma_t} ||g_t||_{t,*}^2 && (7)
\end{aligned}$$

(4) assumes that ψ_t is σ_t-strongly convex with respect to some norm $||\cdot||_t$. (5) uses Hölder's inequality $w^\top v \leq ||w||_t \, ||v||_{t,*}$ where $||\cdot||_{t,*}$ is the dual norm of $||\cdot||_t$ (Appendix M.2.1). (6) minimizes $J(x) = (\frac{\sigma_t}{2} x^2 - \eta_t ||g_t||_* x)$ over $x \in \mathbb{R}$. Finally, (7) uses the convexity of l_t, i.e.,

[1] Since $f(z) = D_{\psi_t}(z, \tilde{w}_{t+1})$ is strictly convex and differentiable, (3) is the unique point $w^\star$ satisfying $\nabla f(w^\star)^\top (w - w^\star) \geq 0$ for all $w \in V$. Likewise, since $h(z) = \eta_t \hat{l}_t(z) + D_{\psi_t}(z, w_t)$ is strictly convex and differentiable, (1) is the unique point $v^\star$ satisfying $\nabla h(v^\star)^\top (w - v^\star) \geq 0$ for all $w \in V$. But $\nabla f(z) = \nabla \psi_t(z) - \nabla \psi_t(\tilde{w}_{t+1}) = \nabla \psi_t(z) - \nabla \psi_t(w_t) + \eta_t g_t = \nabla h(z)$.

$$g_t^\top (w_t - u) \geq l_t(w_t) - l_t(u)$$



Rearranging, we have a per-step regret bound $l_t(w_t) - l_t(u) \leq \frac{1}{\eta_t}\left(D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1})\right) + \frac{\eta_t}{2\sigma_t}\left|\left|g_t\right|\right|_{t,*}^2$. Thus the total regret can be bounded by

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq \sum_{t=1}^{T} \frac{1}{\eta_t}\left(D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1})\right) + \sum_{t=1}^{T} \frac{\eta_t}{2\sigma_t}\left|\left|g_t\right|\right|_{t,*}^2 \tag{8}$$

1.3 Euclidean Analysis

(8) applies to any Bregman divergence. However, for most practical purposes we use squared Euclidean distance $D_{\psi_t}(x, y) = \frac{1}{2}\left|\left|y - x\right|\right|_{A_t}^2$ weighted by a "preconditioner" matrix $A_t \succ 0$.[2] It is induced by $\psi_t(x) = \frac{1}{2}\left|\left|x\right|\right|_{A_t}^2$ which is 1-strongly convex wrt. $\left|\left|\cdot\right|\right|_{A_t}$, thus the second term of (8) becomes $\frac{1}{2}\sum_{t=1}^{T}\eta_t\left|\left|g_t\right|\right|_{A_t^{-1}}^2$. The first term of the bound (8) is now

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{\eta_t}\left(D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1})\right) &= \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\left(\left|\left|w_t - u\right|\right|_{A_t}^2 - \left|\left|w_{t+1} - u\right|\right|_{A_t}^2\right) \\
&= \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t}\left|\left|w_t - u\right|\right|_{A_t}^2 - \frac{1}{\eta_{t-1}}\left|\left|w_t - u\right|\right|_{A_{t-1}}^2\right) - \frac{1}{\eta_T}\left|\left|w_{T+1} - u\right|\right|_{A_T}^2 \qquad (9) \\
&\leq \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t}\left|\left|w_t - u\right|\right|_{A_t}^2 - \frac{1}{\eta_{t-1}}\left|\left|w_t - u\right|\right|_{A_{t-1}}^2\right) \\
&= \frac{1}{2}\sum_{t=1}^{T}(w_t - u)^\top\left(\frac{1}{\eta_t}A_t - \frac{1}{\eta_{t-1}}A_{t-1}\right)(w_t - u) \qquad (10)
\end{aligned}$$

(9) uses the dummy variables $\eta_0 = \infty$ and $A_0 = 0_{d\times d}$. To get a general bound, we can apply $v^\top B v \leq \left|\left|v\right|\right|_2 \left|\left|Bv\right|\right|_2 \leq \mathrm{tr}\left(B\right)\left|\left|v\right|\right|_2^2$ (i.e., the consistency between the matrix spectral norm and the vector l_2 norm) to (10):

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}(w_t - u)^\top\left(\frac{1}{\eta_t}A_t - \frac{1}{\eta_{t-1}}A_{t-1}\right)(w_t - u) &\leq \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t}\mathrm{tr}\left(A_t\right) - \frac{1}{\eta_{t-1}}\mathrm{tr}\left(A_{t-1}\right)\right)\left|\left|w_t - u\right|\right|_2^2 \qquad (11) \\
&\leq \frac{\max_{t=1}^{T}\left|\left|w_t - u\right|\right|_2^2}{2}\underbrace{\sum_{t=1}^{T}\left(\frac{1}{\eta_t}\mathrm{tr}\left(A_t\right) - \frac{1}{\eta_{t-1}}\mathrm{tr}\left(A_{t-1}\right)\right)}_{\text{Telescopes!}} \qquad (12) \\
&= \frac{(\max_{t=1}^{T}\left|\left|w_t - u\right|\right|_2^2)\mathrm{tr}\left(A_T\right)}{2\eta_T} \qquad (13)
\end{aligned}$$

Note that step (12) also requires $\frac{1}{\eta_t}\mathrm{tr}\left(A_t\right) - \frac{1}{\eta_{t-1}}\mathrm{tr}\left(A_{t-1}\right) \geq 0$. To ensure this, we will generally assume that

$$\infty =: \eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots \geq \eta_T > 0 \tag{14}$$
$$0_{d\times d} =: A_0 \prec A_1 \preceq A_2 \preceq \cdots \preceq A_T \tag{15}$$

[2] For $A \succ 0$, $\left|\left|x\right|\right|_A = \sqrt{x^\top A x}$ is a norm on $\mathbb{R}^d$ with $\left|\left|x\right|\right|_{A^{-1}}$ as the dual norm (Appendix M.3). For the sake of simplicity, we will assume that precondioners are positive-definite (e.g., add an infinitesimal value to the diagonal).

If $A_t = A$ across t (i.e., time-invariant preconditioning), we can avoid the lossy inequality (11) since

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}(w_t-u)^\top\left(\frac{1}{\eta_t}A-\frac{1}{\eta_{t-1}}A\right)(w_t-u) &= \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)||w_t-u||_A^2 \\ &\le \frac{\max_{t=1}^T||w_t-u||_A^2}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right) \\ &= \frac{\max_{t=1}^T||w_t-u||_A^2}{2\eta_T} \end{aligned} \tag{16}$$

(16) is d-times sharper than (13) when $A = I_d$. If we further assume that $\eta_t = \eta > 0$ across t (i.e., fixed learning rate), we can make the bound still sharper (recall $\eta_0 = \infty$):

$$\frac{1}{2}\sum_{t=1}^{T}(w_t-u)^\top\left(\frac{1}{\eta_t}A-\frac{1}{\eta_{t-1}}A\right)(w_t-u) = \frac{||w_1-u||_A^2}{2\eta} \tag{17}$$

1.3.1 Algorithm and guarantee

Since $\nabla\psi_t(x) = A_t x$ and $\nabla\psi_t^*(x) = A_t^{-1}x$, we update $\tilde{w}_{t+1} = \nabla\psi_t^*(\nabla\psi_t(w_t) - \eta_t g_t) = w_t - \eta_t A_t^{-1} g_t$. In summary, we start from some initial hypothesis $w_1 \in V \subseteq \mathbb{R}^d$ (with dummy $\eta_0 = \infty$ and $A_0 = 0_{d\times d}$) and for $t = 1, 2, \dots$

- The enemy picks a convex and differentiable loss $l_t : \mathbb{R}^d \to \mathbb{R}$ and we suffer $l_t(w_t) \in \mathbb{R}$.
- We compute the gradient $g_t = \nabla l_t(w_t)$.
- We pick a learning rate $\eta_t \le \eta_{t-1}$ and a preconditioner $A_t \succeq A_{t-1}$.
- We compute $\tilde{w}_{t+1} = w_t - \eta_t A_t^{-1} g_t \in \mathbb{R}^d$.
- We project $w_{t+1} = \arg\min_{w\in V} D_{\psi_t}(w, \tilde{w}_{t+1}) \in V$.

Let $D_A = \max_{t=1}^T ||w_t - u||_A$ and $D_{A,1} = ||w_1 - u||_A$; for the special case $A = I_d$, let $D = \max_{t=1}^T ||w_t - u||_2$ and $D_1 = ||w_1 - u||_2$. After T such steps, we guarantee by (8) that $w_1 \dots w_T \in V$ satisfy

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le \frac{D^2 \operatorname{tr}(A_T)}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t ||g_t||_{A_t^{-1}}^2 \qquad \text{(always)} \tag{18}$$

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le \frac{D_A^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t ||g_t||_{A^{-1}}^2 \qquad \text{(if } A_t = A\text{)} \tag{19}$$

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le \frac{D_{A,1}^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}||g_t||_{A^{-1}}^2 \qquad \text{(if } A_t = A \text{ and } \eta_t = \eta\text{)} \tag{20}$$

In principle, D_A can grow as $O(T)$. This issue is usually addressed by assuming $V \subseteq \mathbb{R}^d$ to have a finite diameter $\Delta = \max_{x,y\in V} ||x - y||_A$ so that $D_A \le \Delta$ is constant in T. But this is not a solution, and in practice $V = \mathbb{R}^d$ almost always (i.e., no projection). Thus we will assume $V = \mathbb{R}^d$ and treat D_A as constant in T.

2 Stochastic Gradient Descent (SGD)

SGD uses no preconditioning (i.e., $A = I_d$) and specifies the update

$$w_{t+1} = w_t - \eta_t g_t \tag{21}$$

corresponding to constant regularization $D_\psi(x, y) = \frac{1}{2}||y - x||_2^2$ in Euclidean space with $\psi(x) = \frac{1}{2}||x||_2^2$. As a special case of mirror descent, it satisfies the regret bound (19) (or (20) if $\eta_t = \eta > 0$ is constant). But the derivation becomes particularly simple, so we give one below. For any t:

$$||w_t - u||_2^2 - ||w_{t+1} - u||_2^2 = ||w_t - u||_2^2 - ||w_t - u - \eta_t g_t||_2^2 = 2\eta_t \underbrace{g_t^\top(w_t - u)}_{\ge l_t(w_t) - l_t(u)} - \eta_t^2 ||g_t||_2^2$$

yielding the per-step bound $l_t(w_t) - l_t(u) \le \frac{1}{2\eta_t}\left(||w_t - u||_2^2 - ||w_{t+1} - u||_2^2\right) + \frac{\eta_t}{2}||g_t||_2^2$ that telescopes and gives us (19) (assuming $\eta_t \le \eta_{t-1}$) and (20). This is so much simpler because we are able to skip the use of the Pythagorean theorem (which holds exactly in this case, check the conditions in Lemma C.3) and other inequalities in (4–6).

2.1 Analysis

Assume a bound on the gradient norm $L \geq \max_{t=1}^{T} ||g_t||_2$ (treated as constant in T). If $\eta_t = \eta$, we have from (20)

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq \frac{D_1^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} ||g_t||_2^2 \leq \frac{D_1^2}{2\eta} + \frac{\eta}{2}L^2 T \tag{22}$$

It is clear that choosing $\eta = \frac{1}{\sqrt{T}}$ makes the bound $O(\sqrt{T})$, but we can explicitly minimize it over η. The minimizer is $\eta^\star = \frac{D_1}{L\sqrt{T}}$ (not practical since D_1 and L are unknown), yielding

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq D_1 L\sqrt{T} \tag{23}$$

If T is unknown, we can use the per-step learning rate $\eta_t = \frac{1}{\sqrt{t}}$. Since it satisfies $\eta_t \leq \eta_{t-1}$, we have from (19)

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq \frac{D^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t ||g_t||_2^2 \leq \frac{D^2}{2}\sqrt{T} + \frac{L^2}{2}\underbrace{\left(\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\right)}_{\leq 2\sqrt{T}} \leq \left(\frac{D^2 + 2L^2}{2}\right)\sqrt{T} \tag{24}$$

where the inequality $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ is a special case of the following fact.

Fact 2.1. Let $x_1 \dots x_T \in \mathbb{R}^d$ be any vectors and let $X_t = \sum_{l=1}^{t} x_l x_l^\top \succeq 0$. Then

$$\sum_{t=1}^{T} x_t^\top X_t^{-1/2} x_t \leq 2\mathrm{tr}\left(X_T^{1/2}\right)$$

If $d = 1$, it can be stated as $\sum_{t=1}^{T}\frac{b_t}{\sqrt{B_t}} \leq 2\sqrt{B_T}$ for any nonnegative $b_1 \dots b_T \geq 0$ where $B_t = \sum_{l=1}^{t} b_l \geq 0$.

Proof. $\mathrm{tr}\left(X^{1/2}\right) \in \mathbb{R}$ is concave in $X \succeq 0$ with gradient $\frac{1}{2}X^{-\top/2}$. Thus $\mathrm{tr}\left(A^{1/2}\right) \leq \mathrm{tr}\left(B^{1/2}\right) + \mathrm{tr}\left(\frac{1}{2}B^{-1/2}(A-B)\right)$ for any $B \succeq A$, or $\mathrm{tr}\left(B^{1/2}\right) - \mathrm{tr}\left(A^{1/2}\right) \geq \frac{1}{2}\mathrm{tr}\left(B^{-1/2}(B-A)\right)$. Since $X_t = X_{t-1} + x_t x_t^\top$, we have $\mathrm{tr}\left(X_t^{1/2}\right) - \mathrm{tr}\left(X_{t-1}^{1/2}\right) \geq \frac{1}{2}\mathrm{tr}\left(X_t^{-1/2} x_t x_t^\top\right) = \frac{1}{2}x_t^\top X_t^{-1/2} x_t$. Summing both sides over t gives the statement. □

2.2 SGD with Polyak Momentum

SGD (21) is vulnerable to gradient noise (e.g., when the Hessian has a large condition number). With Polyak "heavy-ball" momentum, we instead consider

$$v_t = g_t + \mu_t v_{t-1} \tag{25}$$
$$w_{t+1} = w_t - \eta_t v_t \tag{26}$$

where $0 \leq \mu_t < 1$ is the "momentum" (but more like friction) hyperparameter and $v_0 = 0_d$. We will typically assume constant momentum $\mu_t = \mu$. The gradient g_t acts as "acceleration" and only affects the parameter indirectly through "velocity" v_t, which builds up in a direction that has consistent gradient (e.g., $v_4 = g_4 + \mu g_3 + \mu^2 g_2 + \mu^3 g_1$) like a ball rolling down a hill. Note that the accelerated update v_t can be larger than the average gradient. For example, if $g_t = g$ for all t, we have

$$v = \lim_{t\to\infty} v_t = \frac{1}{1-\mu}g \tag{27}$$

(e.g., with the typical value $\mu = 0.95$ we have $v = 20g$). In the OCO setting, additionally assuming that l_t is L-Lipschitz (i.e., $|l_t(w) - l_t(w')| \leq L||w - w'||_2$), we can show that the regret of (25–26) with constant learning rate can achieve the bound (Lemma S.3)

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq D_1 L\sqrt{T}\sqrt{\frac{1+\mu}{1-\mu}}$$

which reverts to the optimal bound for SGD (23) when $\mu = 0$ and blows up as $\mu \to 1^-$.

2.3 SGD with Nesterov Momentum

Nesterov (1983) propose computing the gradient at the hypothetical next location given by the current momentum, which yields the following update[3]

$$x_t = w_t - \eta_t \mu_t v_{t-1} \tag{28}$$
$$g'_t = \nabla l_t(x_t) \tag{29}$$
$$v_t = g'_t + \mu_t v_{t-1} \tag{30}$$
$$w_{t+1} = w_t - \eta_t v_t \tag{31}$$

There have been efforts to relate Polyak and Nesterov momentum in a unified framework; see Appendix F for an example. The gradient may be computed far away from w_t if momentum has built consistently. Nesterov shows that this "peeking" achieves the optimal first-order convergence rate $O(1/T^2)$ in CSO (Section A.1), justifying not looking ahead further.

2.3.1 Double momentum

Naively computing Nesterov would require making two updates per step: one to compute x_t (28), the other to compute w_{t+1} (31). Instead, we typically use the following reparamterization. Assuming constant momentum μ and learning rate η, (28–31) can be written as (Lemma S.10)

$$g'_t = \nabla l_t(x_t)$$
$$v_t = g'_t + \mu_t v_{t-1}$$
$$x_{t+1} = x_t - \eta_t(g'_t + \mu_t v_t)$$

Thus we may update the "fast" weight x_t directly with "double momentum" to perform Nesterov exactly. If $\eta_{t+1} \neq \eta_t$ this is not exact, but for common learning rate schedules (cosine/linear decay) $|\eta_t - \eta_{t+1}| \ll \eta_t$ so implementations usually ignore it.

2.3.2 Double momentum with transform

More advanced opitmizers can be seen as applying some function $\mathbf{transform}_t$ to the final parameter update o_t. Nesterov momentum in this case is (with constant μ)

$$w_{t+1} = w_t - \eta_t \mathbf{transform}_t(g'_t + \mu v_{t-1})$$

(where $g'_t = \nabla l_t(w_t - \eta_t \mu v_{t-1})$ is the future gradient). We have shown that the following double-momentum reparameterization

$$v_t = g_t + \mu_t v_{t-1}$$
$$w_{t+1} = w_t - \eta_t \mathbf{transform}_t(g_t + \mu_t v_t)$$

(where $g_t = \nabla l_t(w_t)$ is the current gradient) is equivalent for SGD (i.e., no transform). The equivalence does not hold in general (e.g., if the transform depends on g_t). Nonetheless, many optimizers perform double momentum in some way and call it "Nesterov", which is functionally more of a variance reduction method.

3 AdaGrad

In (22), SGD uses a constant learning rate $\eta = \frac{1}{\sqrt{T}}$ that ignores the gradient. But we can avoid the lossy second inequality and directly minimize the first bound $\frac{D_1^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T ||g_t||_2^2$ over η. This yields a potentially much larger learning rate $\eta^\star = \frac{D_1}{\sqrt{\sum_{t=1}^T ||g_t||_2^2}} \gg \frac{D_1}{L\sqrt{T}}$ and a tighter bound

$$\sum_{t=1}^T l_t(w_t) - l_t(u) \leq D_1 \sqrt{\sum_{t=1}^T ||g_t||_2^2} \tag{32}$$

[3]The literature often uses the equivalent form:

$$w_{t+1} = y_t - \eta'_t \nabla l_t(y_t)$$
$$y_{t+1} = w_{t+1} + \tau_t(w_{t+1} - w_t)$$

We can show that (28–31) can be mapped to this form for some β_t, y_t, η'_t, and τ_t. Nesterov uses $\tau_t = \frac{t}{t+3}$ for the convergence analysis.

assuming $\sum_{t=1}^{T} ||g_t||_2^2 \ll (\max_{t=1}^{T} ||g_t||_2^2)T$ (i.e., big gradients are outliers). The learning rate requires the knowledge of all gradients $g_1 \dots g_T$, so it can only be set in hindsight. But we can use the *partial sum*:

$$\eta_t = \frac{D}{\sqrt{\sum_{l=1}^{t} ||g_l||_2^2}} \tag{33}$$

Since it satisfies $\eta_t \leq \eta_{t-1}$, in a complete analogy to (24):

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq \frac{D^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t ||g_t||_2^2 \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T} ||g_t||_2^2} + \underbrace{\frac{D}{2}\left(\sum_{t=1}^{T} \frac{||g_t||_2^2}{\sqrt{\sum_{l=1}^{t} ||g_l||_2^2}}\right)}_{\leq 2\sqrt{\sum_{t=1}^{T} ||g_t||_2^2}} \leq \frac{3}{2}\left(D\sqrt{\sum_{t=1}^{T} ||g_t||_2^2}\right) \tag{34}$$

which is only ≈ 1.5 times worse than the oracle bound (32) (assuming $D \approx D_1$). We can scale (33) by $\frac{\sqrt{2}}{2}$ to slightly improve the constant to ≈ 1.4.

Instead of a learning rate, we can similarly find an optimal *preconditioner* in hindsight and work backward. We assume the update $w_{t+1} = w_t + A^{-1}g_t$ (i.e., $\eta = 1$) where $A \succ 0$ has absorbed any fixed learning rate. Denoting $\delta_1 = w_1 - u \in \mathbb{R}^d$ and $O_T = \sum_{t=1}^{T} g_t g_t^\top \succ 0$, we may consider minimizing (20) which is equivalent to

$$A^\star = \underset{A \in \mathbb{R}^{d\times d}:\ A \succeq 0}{\arg\min} \underbrace{\delta_1^\top A \delta_1 + \operatorname{tr}\left(A^{-1} O_T\right)}_{J(A)} \tag{35}$$

This is a proper convex problem. J is bounded below by 0 (both terms are nonnegative). J is convex (the second term is well known to be convex over $A \succ 0$). The feasible set of PSD matrices is closed and convex. Thus an infimum exists. However, no $A^\star \succeq 0$ attains that infimum. $A^\star$ cannot be a boundary point (i.e., has some zero eigenvalues) since then $J(A^\star)$ is undefined. For $A^\star \succ 0$, we must have $\langle \nabla J(A^\star), A - A^\star \rangle_F \geq 0$ for all $A \succeq 0$ which means $\nabla J(A^\star) = 0_{d\times d}$. But this condition is

$$\delta_1 \delta_1^\top = (A^\star)^{-1} O_T (A^\star)^{-1}$$

which is impossible due to a rank mismatch. This implies that while a limit on a series of increasingly degenerate $A \succ 0$ achieves the infimum, the minimizer (35) does not exist.

3.1 Diagonal Preconditioner

AdaGrad (Duchi *et al.*, 2011) dramatically simplifies (35) by constraining the preconditioner to be *diagonal*, i.e., $A = \operatorname{diag}(a_1 \dots a_d)$ where $a_i > 0$. With this restriction, (35) decomposes over dimensions:

$$a_1^\star \dots a_d^\star = \underset{a_1 \dots a_d \geq 0}{\arg\min} \sum_{i=1}^{d} \left(\delta_{1,i}^2 a_i + \frac{1}{a_i}\sum_{t=1}^{T} g_{t,i}^2\right) \tag{36}$$

The objective is convex in each $a_i > 0$. The stationary condition implies the closed form solution

$$a_i^\star = \frac{1}{|\delta_{1,i}|}\sqrt{\sum_{t=1}^{T} g_{t,i}^2} \tag{37}$$

Plugging $A^\star = \operatorname{diag}(a_1^\star \dots a_d^\star)$ in (20), we have the minimized bound

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \leq \sum_{i=1}^{d} |\delta_{1,i}| \sqrt{\sum_{t=1}^{T} g_{t,i}^2} \tag{38}$$

It is instructive to compare this to (32). Letting $\alpha_i = |\delta_{1,i}|$ and $\beta_i = \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$, by Hölder's inequality

$$\sum_{i=1}^{d} |\delta_{1,i}| \sqrt{\sum_{t=1}^{T} g_{t,i}^2} = \alpha^\top \beta \leq ||\alpha||_2 \, ||\beta||_2 = D_1 \sqrt{\sum_{t=1}^{T} ||g_t||_2^2}$$

The equality holds iff $\alpha = \lambda\beta$ for some $\lambda > 0$ (i.e., $A^\star = \lambda I_d$). Otherwise, (38) may be much tighter than (32).

3.1.1 Per-step preconditioner

At step $t \le T$, we again use the partial sum. For instance, if we set $A_{t,i,i} = \frac{1}{D}\sqrt{\sum_{l=1}^{t} g_{l,i}^2}$, we have $A_t \succeq A_{t-1}$ and can straightforwardly use (18) and Fact 2.1 to bound the regret as

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le \frac{3D}{2} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$$

We can also argue for a tighter bound using $A_{t,i,i} = \frac{1}{\delta_i}\sqrt{\sum_{l=1}^{t} g_{l,i}^2}$ where $\delta_i = \max_{t=1}^{T} |w_{t,i} - u_i|$. The idea is to treat $A_{t,i,i}^{-1} =$ (33) as a "learning rate" for each dimension i for which we have the bound (34). We can decompose the bound using the basic fact that "a convex regret is upper bounded by the linearized regret". Formally,

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le \sum_{t=1}^{T} g_t^\top w_t - g_t^\top u = \sum_{i=1}^{d} \left(\sum_{t=1}^{T} g_{t,i} w_{t,i} - g_{t,i} u_i \right) \le \frac{3}{2} \sum_{i=1}^{d} \delta_i \sqrt{\sum_{t=1}^{T} g_{t,i}^2} \tag{39}$$

where the second inequality treats $\sum_{t=1}^{T} g_{t,i} w_{t,i} - g_{t,i} u_i$ as the regret for the 1-dimensional (linear) losses $l_{t,i}(w_{t,i}) = g_{t,i} w_{t,i}$ for each i. This is only ≈ 1.5 times worse than (38) (assuming $|\delta_{1,i}| \approx \delta_i$).

3.2 Full Preconditioner

We can "force" a non-diagonal minimizer in (35) by regularizing the trace:

$$A^\star = \operatorname*{arg\,min}_{A \succ 0} \ \delta_1^\top A \delta_1 + \operatorname{tr}\left(A^{-1} O_T\right) \approx \operatorname*{arg\,min}_{A \succ 0:\ \operatorname{tr}(A) \le c} \operatorname{tr}\left(A^{-1} O_T\right) = \frac{c}{\operatorname{tr}(O_T^{1/2})} O_T^{1/2} \tag{40}$$

(see this note for a derivation of the closed-form solution). Plug in $A^\star = O_T^{1/2}$ in (18) with $\eta_t = D$ to have

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le D \operatorname{tr}\left(O_T^{1/2}\right) \tag{41}$$

Compare this with plugging in the diagonal counterpart $A_{i,i}^\star = \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$ in (18) which yields

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le D \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2} \tag{42}$$

Though they look similar, (41) is smaller than (42) unless O_T is diagonal (Lemma S.1). Intuitively, $O_T^{1/2}$ exploits the interplay between dimensions while $\sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$ does not. In particular, (41) is $O(\sqrt{T})$ (since (42) is).

3.2.1 Per-step preconditioner

At step $t \le T$, let $O_t = \sum_{l=1}^{t} g_l g_l^\top$ and use $A_t = O_t^{1/2}$. Assuming the constant learning rate $\eta_t = D$, (18) becomes

$$\sum_{t=1}^{T} l_t(w_t) - l_t(u) \le \frac{D \operatorname{tr}\left(O_T^{1/2}\right)}{2} + \frac{D}{2} \sum_{t=1}^{T} g_t^\top O_t^{-1/2} g_t \le \frac{D \operatorname{tr}\left(O_T^{1/2}\right)}{2} + D \operatorname{tr}\left(O_T^{1/2}\right) = \frac{3D}{2} \operatorname{tr}\left(O_T^{1/2}\right) \tag{43}$$

where the second inequality uses Fact 2.1. Again, we conclude that the regret bound is only 1.5 times worse when using a per-step preconditioner (i.e., compared to (41)).

3.3 AdaGrad in Practice

The full AdaGrad preconditioner $A_t = (\sum_{l=1}^{t} g_l g_l^\top)^{1/2} \in \mathbb{R}^{d \times d}$ is unfortunately impractical for any large d, so we use the diagonal preconditioner where $A_{t,i,i} = \sqrt{\sum_{l=1}^{t} g_{l,i}^2}$. Then the update $w_{t+1} = w_t + \eta_t A_t^{-1} g_t$ is equivalent to per-parameter adaptive learning rates:

$$w_{t+1,i} = w_{t,i} - \frac{\eta_t}{\sqrt{\sum_{l=1}^{t} g_{l,i}^2}} g_{t,i} \tag{44}$$

where the learning rate shrinks based on how heavily the parameter has been updated in the past. Note that $w_{2,i} = w_{1,i} - \eta_1$ and the magnitude of the update is always at most η_t.

4 Adam

AdaGrad has inspired a whole class of per-parameter adaptive updates. One practical issue of AdaGrad is that the update can only become smaller throughout training because the denominator in (44) can only become larger. This is fine for convex problems where there is only one local optimum, but we may want to allow the update to jump back in size for nonconvex problems. One way to address this issue is by "forgetting" the far past. We may only use the past $K < t$ steps at step t to compute the denominator. Better, we may use an exponential moving average (EMA):

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

where $v_0 = 0_d$ and $\beta_2 \in [0, 1)$ is a coefficient (i.e., how much to remember). If we view g_t as iid random variables, we have $\mathbf{E}[v_t] = (1-\beta_2^t)\mathbf{E}[g_t^2]$ (Lemma S.11) which converges to the true second moment as $t \to \infty$. While at it, we can use momentum for the gradient itself which is well known to help in making SGD more stable:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

for $m_0 = 0_d$ and $\beta_1 \in [0, 1)$. Again, $\mathbf{E}[m_t] = (1-\beta_1^t)\mathbf{E}[g_t]$ as $t \to \infty$. Replacing g_t and $\sum_{l=1}^t g_{l,i}^2$ in (44) with m_t and v_t, we have **RMSProp** with momentum (Tieleman *et al.*, 2012; Graves, 2013):

$$w_{t+1} = w_t - \eta_t \frac{m_t}{\sqrt{v_t}} \tag{45}$$

Adam (Kingma and Ba, 2014) observes that since β_1, β_2 are typically close to 1, the initial updates will be small until they gain some momentum. But since

$$\mathbf{E}[g_t] = \frac{1}{1-\beta_1^t}\mathbf{E}\left[m_t\right] \qquad\qquad \mathbf{E}[g_t^2] = \frac{1}{1-\beta_2^t}\mathbf{E}\left[v_t\right]$$

we can use $\bar{m}_t = \frac{1}{1-\beta_1^t} m_t$ and $\bar{v}_t = \frac{1}{1-\beta_2^t} v_t$ to correct the bias so that $\mathbf{E}[\bar{m}_t] = \mathbf{E}[g_t]$ and $\mathbf{E}[\bar{v}_t] = \mathbf{E}[g_t^2]$. This yields the Adam update

$$w_{t+1} = w_t - \eta_t \frac{\bar{m}_t}{\sqrt{\bar{v}_t}} \tag{46}$$

Using the square root of the second moment $\mathbf{E}[g_t^2]$ instead of the sum $\sum_t g_t^2$ is a fundamental departure from AdaGrad. In this case, the preconditioner can be seen as a flawed approximation of the Hessian/Fisher matrix, relating Adam to Netwon's method and natural gradient descent (Section 6.1).

4.1 Scale Invariance

An important property of any AdaGrad-style update like Adam and RMSProp is scale invariance: the gradient can be scaled by an arbitrary constant per dimension without changing the update. More specifically, we can multiply all gradients g_t elementwise by some $c \in \mathbb{R}^d$ in Adam and have

$$w_{t+1} = w_t - \eta_t \frac{\operatorname{diag}(c)\,\bar{m}_t}{\sqrt{\operatorname{diag}(c)^2\,\bar{v}_t}} = w_t - \eta_t \frac{\bar{m}_t}{\sqrt{\bar{v}_t}} \tag{47}$$

Intuitively, scale invariance makes training deep networks easier. The gradient of a linear function $c^\top w$ with respect to w (weight) is c (activation), which may blow up or shrivel in top layers. With vanilla SGD, top layers may receive huge or tiny updates unless we carefully adjust initialization and layerwise learning rates. With scale-invariant methods like Adam, pure rescalings of activations (and hence gradients) do not change the normalized update, so learning is much less sensitive to gradient scale across layers.

4.1.1 Epsilon smoothing

During training, $\bar{v}_t \approx g_t^2$ may become too small for the employed machine precision and underflow to 0, causing the update $\bar{m}_t/\sqrt{\bar{v}_t}$ to be undefined. In practice, Adam is implemented with smoothing by $\epsilon > 0$:

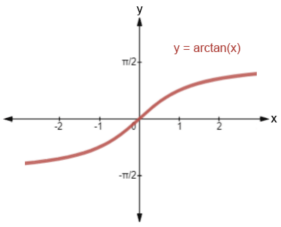
$$w_{t+1} = w_t - \frac{\eta_t}{\sqrt{\bar{v}_t} + \epsilon}\bar{m}_t$$

(sometimes applied inside square-root). If $|g_t| \gg \epsilon$, the smoothing is largely harmless and empirically effective; scale invariance (47) is also approximately preserved (Zhuang *et al.*, 2022). Otherwise, it clearly interferes with the correctness of the update and breaks scale invariance. This becomes an issue especially when the model is large, since $|g_t| = O(1/d)$ in typical inverse-width initialization, necessiating epsilon tuning.

4.1.2 Adam-atan2

One alternative to epsilon smoothing is to "offload" the handling of division by zero to an existing function. Any function of form $f(x, y) \approx x/y$ for $x/y \lesssim 1$ whose implementation robustly handles the case $y \approx 0$ can be used as a drop-in replacement. Everett *et al.* (2024) note that

$$\text{atan2}(\bar{m}_t, \sqrt{\bar{v}_t}) := \arctan\left(\frac{\bar{m}_t}{\sqrt{\bar{v}_t}}\right) \approx \frac{\bar{m}_t}{\sqrt{\bar{v}_t}} \quad \text{for } |\bar{m}_t| \ll \bar{v}_t$$



fits the bill with the possible benefit of "clipping" large update size (which is done explicitly in some variants of Adam, e.g., see 3) to at most $\pi/2 \approx 1.57$.[4] Under the hood, a well-oiled implementation of $\text{atan2}(x, y)$ computes $\arctan(x/y)$ by numerical approximation and is safely defined in edge cases (e.g., $\text{atan2}(0, 0) := 0$). We emphasize that this approach is not a solution to the numerical instability problem and depends on the battle-tested robustness of atan2 implementation.

4.2 Convergence

In the proof of AdaGrad's convergence, we use the fact that the learning rate is nonincreasing. This is no longer true in momentum-based updates like RMSProp and Adam. Reddi *et al.* (2019) demonstrate that Adam does not converge on a convex problem (i.e., it has a linear regret) and propose to enforce a nonincreasing learning rate by maintaining the elementwise max for the second moment (AMSGrad). They use the formulation without bias correction: $v_t^{\max} = \max(v_{t-1}^{\max}, v_t)$ where $v_0^{\max} = 0_d$. AMSGrad now has convergence guarantees and is able to converge on synthetic examples that vanilla Adam fails to converge in (Figure 1 in their paper). In practice, however,

1. AMSGrad does not seem to make a whole lot of difference in downstream performance (see this blog).

2. With bias correction (no guidance from the paper), the max needs to be applied after, not before, to achieve a nonincreasing learning rate. However, several implementations (e.g., PyTorch) apply the max *before* the correction.

That said, the wrong order in common implementations matters little after a few hundred steps since $1 - \beta_2^t \approx 1$. But we will do the correct thing and apply the max after bias correction (Figure 1).

4.3 Nesterov Momentum

The bias-corrected first moment takes the form

$$\bar{m}_t = \frac{1}{1-\beta_1^t} m_t = \frac{\beta_1}{1-\beta_1^t} m_{t-1} + \frac{1-\beta_1}{1-\beta_1^t} g_t$$

Dozat (2016) propose to replace m_{t-1} in the last expression with m_t to implement the double momentum trick in Nesterov. Accounting for the fact that m_t is one step ahead, this yields

$$\bar{m}_t = \frac{\beta_1}{1-\beta_1^{t+1}} m_t + \frac{1-\beta_1}{1-\beta_1^t} g_t$$

Unlike in SGD, this trick is not equivalent to the true Nesterov (i.e., using the future gradient $g_t' = \nabla l_t(w_t - \eta_t \beta_1 \bar{m}_{t-1})$ given by the current first moment).

[4] They introduce two knobs a, b to keep $a \cdot \text{atan2}(\bar{m}_t, b \cdot \sqrt{\bar{v}_t}) = \bar{m}_t / \sqrt{\bar{v}_t}$ for two regimes: $|\bar{m}_t| \ll \sqrt{\bar{v}_t}$ and $|\bar{m}_t| \approx \sqrt{\bar{v}_t}$ (they argue heuristically that the "typical" regime has $|\bar{m}_t| \lesssim \sqrt{\bar{v}_t}$ since $\bar{m}_t \approx g_t$ and $\bar{v}_t \approx g_t^2$ with $|\mathbf{E}[g_t]| \le \sqrt{\mathbf{E}[g_t^2]}$ by Jensen). The former requires $a = b$ while the latter requires $a = 1/\arctan(1/b)$, converging as $b \to \infty$ but different otherwise. They end up choosing $a = b = 1$, which means the update size may be $\approx 0.8\times$ smaller in the latter regime. But the difference is minor and easily absorbed in learning rate tuning.

4.4 Weight Decay

Hanson and Pratt (1988) originally proposed weight decay $w \leftarrow (1-\lambda)w$ as a way of regularizing the model size independently of the loss. In SGD, it coincides with l_2 regularization. Even here, there is an important caveat: the decay factor must be coupled with the learning rate.

$$w_{t+1} = \underbrace{w_t - \eta_t \nabla \left(l_t(w_t) - \frac{\lambda'}{2} ||w_t||_2^2 \right)}_{\text{SGD with } l_2 \text{ regularization}} = \underbrace{w_t - \eta_t \nabla l_t(w_t) - \eta_t \lambda' w_t}_{\text{SGD with decay factor } \lambda = \eta_t \lambda'}$$

With pre-conditioning they do not coincide.

$$w_{t+1} = w_t - \eta_t A^{-1} \nabla \left(l_t(w_t) - \frac{\lambda'}{2} ||w_t||_2^2 \right) = w_t - \eta_t A^{-1} \nabla l_t(w_t) - \lambda' \eta_t A^{-1} w_t$$

where the last term is not equal to λw_t for any $\lambda > 0$ unless $A = cI_d$ is spherical. Loshchilov and Hutter (2017) thus propose to perform explicit weight decay on top of Adam, denoted as **AdamW**. The original AdamW paper describes coupled weight decay, which is presumably the reason that standard libraries multiply the decay factor with the learning rate (e.g., PyTorch). However, Wortsman *et al.* (2024) find that fully decoupling the learning rate and the decay factor reduces the sensitivity of training to the choice of learning rate. Specifically, they use

$$w_{t+1} = w_t - s_t \left(\eta_{\max} \frac{\bar{m}_t}{\sqrt{\bar{v}_t}} + \lambda_{\max} w_t \right)$$

where $\eta_{\max}$ and $\lambda_{\max}$ are the maximum learning rate and decay factor, and $s_t \in [0, 1]$ is a schedule multiplier. The schedule typically "warms up" for T_{warmup} steps to 1, then "cools down" to some small final value. Thus the update has the form $w_{t+1} = w_t - \eta_t \frac{\bar{m}_t}{\sqrt{\bar{v}_t}} - \lambda_t w_t$ with the stepwise $\eta_t = s_t \eta_{\max}$ and $\lambda_t = s_t \lambda_{\max}$. In contrast, coupled weight decay has the form $w_{t+1} = w_t - \eta_t (\frac{\bar{m}_t}{\sqrt{\bar{v}_t}} - \lambda_{\max} w_t)$.

4.4.1 Scaling relation to learning rate

Let $w_{t+1} = (1-\lambda)w_t - \eta o_t$ with constant $0 < \lambda < 1$ and $\eta > 0$, where the updates o_t are iid with mean 0 and variance $\sigma^2 > 0$.[5] Then $w_t = -\eta \sum_i (1-\lambda)^i o_i$ has the stationary size

$$|w_\infty| := \lim_{t \to \infty} \sqrt{\mathbf{E}[w_t^2]} = \frac{\eta \sigma}{\sqrt{\lambda(2-\lambda)}} \approx \frac{\eta \sigma}{\sqrt{2\lambda}} \tag{48}$$

safely assuming $\lambda \approx 0$ ($\lambda \lesssim 0.01$ even in aggressive cases). With coupled WD $w_{t+1} = (1 - \lambda\eta)w_t - \eta o_t$, this becomes

$$|w_\infty| \approx \sqrt{\frac{\eta}{2\lambda}} \sigma \tag{49}$$

$|w_\infty|$ diverges when $\lambda = 0$, so the analysis is not useful for comparing WD with "no WD".[6] However, it guides us how to scale λ in relation to η (and σ). E.g., under decoupled WD, if we want to maintain $|w_\infty|$ with $\eta \mapsto c\eta$, we should analogously map $\lambda \mapsto c^2 \lambda$; if we want $|w_\infty| = a$ for some target size $a \in \mathbb{R}$, we should set $\lambda = \Theta((\eta/a)^2)$.

4.5 Asymptotic Update Size

We assume $d = 1$ and analyze $\sqrt{\mathbf{E}[o_t^2]} \in \mathbb{R}$ as $t \to \infty$ as the stationary size of Adam $o_t = \frac{\bar{m}_t}{\sqrt{\bar{v}_t}}$ (Appendix K). Assuming $\mathbf{E}[g_t] = 0$, by the usual properties of EMA[7]

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \xrightarrow{d} M \qquad \mathbf{E}[M] = 0 \qquad \text{Var}(M) = \mathbf{E}[g_t^2] \frac{1-\beta_1}{1+\beta_1}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \xrightarrow{d} V \qquad \mathbf{E}[V] = \mathbf{E}[g_t^2] \qquad \text{Var}(V) = \text{Var}\left(g_t^2\right) \frac{1-\beta_2}{1+\beta_2}$$

[5] The iid assumption on o_t holds in momentumless Adam with $\sigma^2 = 1$ (assuming the gradients are iid and centered). But it is generally false since $o_t = \bar{m}_t / \sqrt{\bar{v}_t}$ is correlated across steps through EMAs (though it is possible that the correlation is mild).

[6] Intuitively, the noise wins in the infinite horizon if there is no contraction: the variance diverges like a random walk.

[7] Shortcut: at stationary state, an EMA has the form $Z = \beta Z + (1-\beta)X$, which implies $\mathbf{E}[Z] = \mathbf{E}[X]$ and $\text{Var}(Z) = \frac{(1-\beta)^2}{1-\beta^2} \text{Var}(X)$. See Lemma S.11 for details.

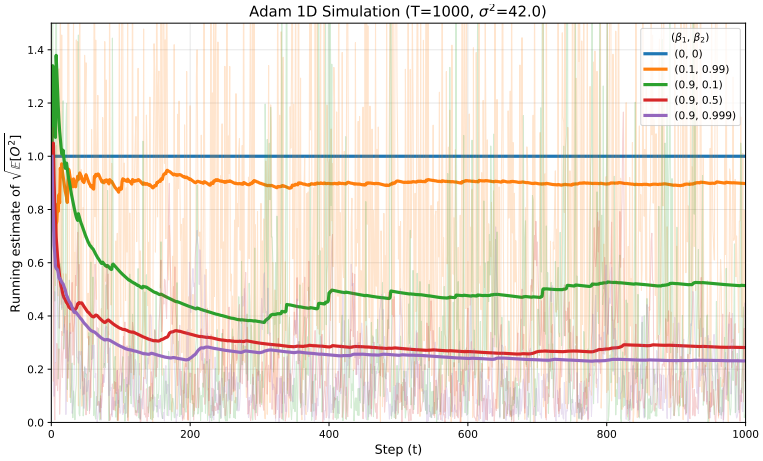
Thus $(\bar{m}_t, \bar{v}_t) \xrightarrow{d} (M, V)$ and $o_t \xrightarrow{d} O := \frac{M}{\sqrt{V}}$ as well in the infinite horizon. We can show (Lemma S.16)

$$\sqrt{\mathbf{E}[O^2]} = \sqrt{\frac{1-\beta_1}{1+\beta_1}} + O(1-\beta_2) \tag{50}$$

where the asymptotic term is in $\beta_2 \to 1^-$. A quick "hack" to see this is observing that $\mathrm{Var}\,(V)$ tends to zero as β_2 nears one, allowing us to treat $V = \mathbf{E}[g_t^2]$ as constant and have

$$\sqrt{\mathbf{E}[O^2]} = \sqrt{\mathbf{E}\left[\frac{M^2}{V}\right]} \approx \sqrt{\frac{\mathbf{E}\left[M^2\right]}{\mathbf{E}[g_t^2]}} = \sqrt{\frac{1-\beta_1}{1+\beta_1}} \tag{51}$$

(50) shows that the effect of β_2 is linear and thus mild compared to the strong dependence on β_1, so (51) remains a good rule of thumb. Below we plot running estimates of $\sqrt{\mathbf{E}[O^2]}$ drawing $g_t \sim \mathcal{N}(0, \sigma^2 = 42)$:[8]



Caveat: note that the instantaneous update $|o_t|$ itself keeps fluctuating and does not converge; what converges is the distribution of o_t and the "typical" size $\sqrt{\mathbf{E}[o_t^2]} \approx |o_t|$.

4.6 Full Algorithm

The full AdamW algorithm is given in Figure 1. The hyperparameters affect each other and need to be tuned jointly for the given model and dataset (e.g., a longer warmup allows for a larger value of effective $\eta_{\max}$). There are many techniques specifically designed for large-scale training. One example is "annealing", in which the final phase of training is performed on very high quality data with a schedule that linearly decays to 0. An average of the weights during annealing is used as the final model (Dubey *et al.*, 2024).

4.7 Adafactor

A practical issue with Adam is that it requires maintaining the first/second gradient moments $m, v \in \mathbb{R}^d$. For instance, if $w \in \mathbb{R}^d$ are in `bfloat16`, maintaining m, v in `float32` increases the memory requirement for optimization from $4d$ to $12d$ bytes (excluding other overheads in backpropagation). **Adafactor** (Shazeer and Stern, 2018) assumes that weights are organized as matrices $W \in \mathbb{R}^{m \times n}$ (e.g., layers) and uses a low-rank approximation of the corresponding second moment $V \in \mathbb{R}^{m \times n}$. If there are $\ll d$ weight matrices, this effectively makes the memory overhead $O(1)$. The usual Adam update has the form (for matrix weights)

$$W_{t+1} = W_t - \eta \frac{M_t}{\sqrt{V_t}}$$

where $M_t \approx \mathbf{E}[G_t]$ and $V_t \approx \mathbf{E}[G_t^2] = G^2$ for the stochastic gradient $G_t \in \mathbb{R}^{m \times n}$. Adafactor instead proposes to perform

$$W_{t+1} = W_t - \eta \frac{M_t}{\sqrt{A_t B_t}}$$

[8] Under mild assumptions, the Markov process $o_t \to O$ is ergodic so that $1/T \sum_{t=1}^{T} o_t^2 \to \mathbf{E}[O^2]$ as $T \to \infty$.

Input:

- Initial parameter value $w_1 \in \mathbb{R}^d$
- Loss functions $l_1, l_2, \ldots, l_T : \mathbb{R}^d \to \mathbb{R}$ (l_t corresponds to the loss on the t-th minibatch)
- Schedule $s_1, s_2, \ldots, s_T \in [0, 1]$
- Maximum learning rate $\eta_{\max} > 0$; maximum weight decay factor $0 \leq \lambda_{\max} < 1$
- Momentum coefficients (β_1, β_2); smoothing coefficient $\epsilon \geq 0$; flag for AMSGrad, Nesterov, DecoupledWD

1. Initialize the first/second momentum estimates $(m_0, v_0, \bar{v}_0^{\max}) \leftarrow (0_d, 0_d, 0_d)$.
2. For $t = 1 \ldots T$:
   (a) Do forward/backward and compute the gradient $g_t \leftarrow \nabla l_t(w_t)$.
   (b) Compute the bias-corrected EMA estimates for the gradient and squared gradient:

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t \qquad v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$\bar{m}_t \leftarrow \begin{cases} \frac{1}{1-\beta_1^t} m_t & \text{(default)} \\ \frac{\beta_1}{1-\beta_1^{t+1}} m_t + \frac{1-\beta_1}{1-\beta_1^t} g_t & \text{(if Nesterov)} \end{cases} \qquad \bar{v}_t \leftarrow \frac{1}{1-\beta_2^t} v_t \qquad \bar{v}_t^{\max} = \max(\bar{v}_{t-1}^{\max}, \bar{v}_t)$$

   (c) If AMSGrad, overwrite $\bar{v}_t \leftarrow \bar{v}_t^{\max}$.
   (d) Get the per-step learning rate and weight decay

$$\eta_t = s_t \eta_{\max} \qquad \lambda_t = \begin{cases} \eta_t \lambda_{\max} & \text{(default)} \\ s_t \lambda_{\max} & \text{(if DecoupledWD)} \end{cases}$$

   (e) Compute the per-parameter update for $i = 1 \ldots d$:

$$w_{t+1,i} \leftarrow (1-\lambda_t) w_{t,i} - \frac{\eta_t}{\sqrt{\bar{v}_{t,i}} + \epsilon} \bar{m}_{t,i}$$

3. Return $w_{T+1} \in \mathbb{R}^d$

Figure 1: The full AdamW algorithm.

where $A_t \in \mathbb{R}^{m \times r}$ and $B_t \in \mathbb{R}^{r \times n}$ are low-rank matrices such that $\mathbf{E}[A_t B_t] \approx G^2$. Practical considerations impose certain constraints: (1) A_t, B_t need to be updatable in an online fashion, (2) they are (ideally) strictly positive since we will divide by their square roots. This makes SVD difficult to use (though it yields an optimal solution in Frobenius norm) since it does not decompose over matrix additions and can be negative. The problem is more naturally approached as nonnegative matrix factorization (NMF) (Appendix L). It is well known that the following rank-1 NMF objective

$$a^\star, b^\star \in \operatorname*{arg\,min}_{a \in \mathbb{R}^m_{\geq 0}, b \in \mathbb{R}^n_{\geq 0}} \operatorname{IDiv}(G^2, ab^\top)$$

has the solution space of $a^\star (b^\star)^\top = \frac{G^2 1_n 1_m^\top G^2}{1_m^\top G^2 1_n}$ (e.g., $a^\star = G^2 1_n$ and $b^\star = \frac{(G^2)^\top 1_m}{1_m^\top a^\star}$). To derive an online update, Adafactor maintains the EMA (with $a_0 = 0_m$ and $s_0 = 0_n$):

$$a_t = \beta_2 a_{t-1} + (1-\beta_2) G_t^2 1_n \qquad \bar{a}_t = \frac{1}{1-\beta_2^t} a_t \qquad \widehat{V}_t = \frac{\bar{a}_t \bar{s}_t^\top}{1_m^\top \bar{a}_t} = \left(\frac{1}{1-\beta_2^t}\right) \frac{a_t s_t^\top}{1_m^\top a_t} \tag{52}$$

$$s_t = \beta_2 s_{t-1} + (1-\beta_2)(G_t^2)^\top 1_m \qquad \bar{s}_t = \frac{1}{1-\beta_2^t} s_t$$

and uses $\widehat{V}_t$ to approximate the second moment.[9] While the rank-1 constraint can be limiting, it is easy to derive a rank-r generalization of Adafactor using EM (Appendix L.2).

[9] Note that this is a biased estimator of the optimal rank-1 decomposition since a_t and s_t are correlated. That is, $\mathbf{E}[\widehat{V}_t] = \mathbf{E}[\frac{\bar{a}_t \bar{s}_t^\top}{1_m^\top \bar{a}_t}] \neq \frac{\mathbf{E}[\bar{a}_t]\mathbf{E}[\bar{s}_t]^\top}{1_m^\top \mathbf{E}[\bar{a}_t]} = \frac{G^2 1_n 1_m^\top G^2}{1_m^\top G^2 1_n} = a^\star (b^\star)^\top$.

4.7.1 Bells and whistles

While rank-1 approximation of the second moment is the "main feature" of Adafactor, it introduces several other modifications to Adam and results in an update that more resembles the following:

$$W_{t+1} = W_t - \eta \times \mathrm{RMS}(W_t) \times \mathbf{RMSClip}\left(\frac{G_t}{\sqrt{\widehat{V}_t}}\right)$$

(η is a global learning rate schedule). The changes include

1. (Main) The use of rank-1 $\widehat{V}_t$ in (52) to cut down memory
2. Not maintaining the first moment (i.e., $\beta_1 = 0$) and using the raw gradient G_t to further cut down memory
3. Clipping the layerwise update $O_t \mapsto \frac{O_t}{\max(1, \mathrm{RMS}(O_t))}$ to make its RMS ≤ 1
4. Parameter scaling: scale the effective learning rate by $\mathrm{RMS}(W_t)$, so that after update clipping the RMS of the parameter change is layerwise scaled to be $\le \eta \mathrm{RMS}(W_t)$
5. Replacing bias correction with $\beta_2 := 1 - t^{-c}$ scheduling

Confusingly, the literature uses the name "Adafactor" to refer to any combination of these changes, even when it does not involve any factorization! For instance, in the context of scale-invariant parameterizations, Adafactor is mainly 4, i.e., a way to scale the learning rate adaptively to match the weight RMS (Everett *et al.*, 2024).

5 Shampoo

AdaGrad shows that the best we can do is $w_{t+1} = w_t - \eta O_t^{-1/2} g_t$ where $O_t = \sum_{l=1}^t g_l g_l^\top \in \mathbb{R}^{d \times d}$. But this incurs $O(d^3)$ compute overhead (i.e., to invert a $d \times d$ matrix). Instead of resorting to a diagonal appoximation, **Shampoo** (Gupta *et al.*, 2018) proposes a clever middle ground by assuming the hypothesis space $\mathbb{R}^{m \times n}$ of matrices. At step t, we propose $W_t \in \mathbb{R}^{m \times n}$ and receive a loss $l_t(W_t) \in \mathbb{R}$ where $l_t : \mathbb{R}^{m \times n} \to \mathbb{R}$ is convex and differentiable. Let $G_t = \nabla l_t(W_t) \in \mathbb{R}^{m \times n}$ denote the per-step gradient. Shampoo prescribes

$$W_{t+1} = W_t - \eta \underbrace{L_t^{-1/4}}_{m \times m} \underbrace{G_t}_{m \times n} \underbrace{R_t^{-1/4}}_{n \times n} \qquad L_t = \sum_{l=1}^t G_l G_l^\top \qquad R_t = \sum_{l=1}^t G_l^\top G_l \tag{53}$$

The compute overhead is now $O(m^3 + n^3)$ which is much smaller than the full conditioning overhead $O(m^3 n^3)$. For analysis, we can convert (53) to an equivalent standard form by the usual properties of Kronecker product (Appendix G),

$$w_{t+1} = w_t - \eta \underbrace{\left(L_t^{1/4} \otimes R_t^{1/4}\right)^{-1}}_{mn \times mn} \underbrace{g_t}_{mn \times 1} \tag{54}$$

where $g_t = \overline{\mathrm{vec}}(G_t)$ and $w_t = \overline{\mathrm{vec}}(W_t)$. Thus Shampoo is "just" Euclidean mirror descent with the per-step preconditioner $A_t = L_t^{1/4} \otimes R_t^{1/4}$. Since $L_t \succeq L_{t-1}$ and $R_t \succeq R_{t-1}$, we also have $A_t \succeq A_{t-1}$ (see (107)). The obvious intuition is that $A_t \approx O_t^{1/2}$. In particular, we can show that

$$O_t^{1/2} \preceq \sqrt{r}(L_t^{1/4} \otimes R_t^{1/4}) = \sqrt{r} A_t \tag{55}$$

where $r = \max_t \operatorname{rank}(G_t)$.[10] Since the vectorized losses $l_t : \mathbb{R}^{mn} \to \mathbb{R}$ (trivially) remain convex and differentiable and $A_t \succeq A_{t-1}$, we can use (18) to bound the regret:

$$\begin{aligned}
\sum_{t=1}^{T} l_t(w_t) - l_t(u) &\le \frac{D^2 \operatorname{tr}(A_T)}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} g_t^\top A_t^{-1} g_t \\
&\le \frac{D^2 \operatorname{tr}(A_T)}{2\eta} + \frac{\eta\sqrt{r}}{2}\sum_{t=1}^{T} g_t^\top O_t^{-1/2} g_t && \text{(since } A_t^{-1} \preceq \sqrt{r} G^{-1/2} \text{ by (55))} \\
&\le \frac{D^2 \operatorname{tr}(A_T)}{2\eta} + \eta\sqrt{r}\operatorname{tr}\left(O_T^{1/2}\right) && \text{(Fact 2.1)} \\
&\le \frac{D^2 \operatorname{tr}(A_T)}{2\eta} + \eta r \operatorname{tr}(A_T) && \text{(using (55) again)} \\
&= D\sqrt{2r}\,\operatorname{tr}\left(L_T^{1/4}\right)\operatorname{tr}\left(R_T^{1/4}\right) && \left(\text{using } \eta = \frac{D}{\sqrt{2r}}\right)
\end{aligned}$$

We can show that $\operatorname{tr}(L_T^{1/4}) = O(T^{1/4})$ and $\operatorname{tr}(R_T^{1/4}) = O(T^{1/4})$, thus the bound is $O(\sqrt{T})$.

5.1 Shampoo with EMA

Shampoo is derived as an approximation to the AdaGrad preconditioner (55) and therefore uses the sum of the gradient outer products (i.e., $L_t = \sum_{l \le t} G_l G_l^\top$ and $R_t = \sum_{l \le t} G_l^\top G_l$). As with RMSProp/Adam, in practice we benefit from replacing it with a running estimate of the expected value $L = \mathbf{E}[G_t G_t^\top]$ and $R = \mathbf{E}[G_t^\top G_t]$, e.g., bias-corrected EMA, so that the update is not made monotonically smaller (Shi *et al.*, 2023). We can then view Shampoo as

$$W_{t+1} = W_t - \eta L^{-1/4} G_t R^{-1/4} \qquad \Leftrightarrow \qquad w_{t+1} = w_t - \eta\big(\underbrace{L^{1/4} \otimes R^{1/4}}_{A_{\text{shampoo}}}\big)^{-1} g_t$$

By adapting (55), we can easily show

$$I_{\text{emp}}^{1/2} \preceq \sqrt{r} L^{1/4} \otimes R^{1/4} = \sqrt{r} A_{\text{shampoo}} \tag{56}$$

So as in RMSProp/Adam, Shampoo with EMA can be motivated as approximating $A_{\text{shampoo}} \approx I_{\text{emp}}^{1/2} \approx I_{\text{fisher}} \approx H$. Instead of using the bound (56) for approximation, Morwani *et al.* (2024) directly approximate $I_{\text{emp}}^{1/2}$ with one round of power iteration and derive the *squared* preconditioner $A_{\text{shampoo}}^2 = L^{1/2} \otimes R^{1/2}$ (Section 6.2).

6 The Hessian View

Let $w \in \mathbb{R}^d$ denote the weight of (some layer of) a neural network. Given a labeled input $(x, y) \in \mathcal{X} \times \{1 \dots K\}$, let $f_w(x) \in \mathbb{R}^K$ denote the final logits and $L(f_w(x), y) = -\log p_{f_w(x)}(y) \in \mathbb{R}$ the cross-entropy loss. Let $g_{x,y}(w) = \frac{\partial L(f_w(x),y)}{\partial w} \in \mathbb{R}^d$ and $H_{x,y}(w) = \frac{\partial^2 L(f_w(x),y)}{\partial w^2} \in \mathbb{R}^{d \times d}$ the gradient and the Hessian on (x, y). Let $g(w) = \mathbf{E}[g_{x,y}(w)]$ and $H(w) = \mathbf{E}[H_{x,y}(w)]$ denote the expected gradient and the expected Hessian where $(x, y) \sim \mathbf{pop}$. We will assume that the fastest way to converge to a critical point of the loss $J(w) = E_{(x,y)\sim\mathbf{pop}}[L(f_w(x), y)]$ is Newton's method:

$$w \leftarrow w - \eta H(w)^{-1} g(w) \tag{57}$$

6.1 The Hessian View of Adam

Can we estimate the Hessian using only the gradient? We have

$$H(w) \overset{(113)}{\approx} H_{\text{GN}}(w) \overset{(H.1)}{=} I(w) = \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop} \\ \hat{y}\sim f_w(x)}} \left[g_{x,\hat{y}}(w) g_{x,\hat{y}}(w)^\top\right] = \mathop{\operatorname{Cov}}_{\substack{x\sim\mathbf{pop} \\ \hat{y}\sim f_w(x)}} \left(g_{x,\hat{y}}(w)\right) \tag{58}$$

[10] From Lemma G.3, we have $O_t = \sum_{l=1}^t g_l g_l^\top \preceq r \sum_{l=1}^t (G_l G_l^\top) \otimes I_n = r L_t \otimes I_n$ and similarly $O_t \preceq r I_m \otimes R_t$. Using Fact G.4, we get $O_t \preceq r(L_t \otimes I_n)^{1/2}(I_m \otimes R_t)^{1/2} = r L_t^{1/2} \otimes R_t^{1/2}$ and also $O_t^{1/2} \preceq \sqrt{r}(L_t^{1/4} \otimes R_t^{1/4})$.

where $H_{\text{GN}}(w)$ is the Gauss-Newton component and $I(w)$ is the Fisher matrix. The last equality is the well-known covariance characterization of the Fisher, which follows since $\mathbf{E}[g_{x,\hat{y}}(w)] = 0_d$ for $\hat{y} \sim f_w(x)$. The Fisher matrix is difficult to estimate because it is an expectation over the model's distribution (a moving target). A standard approach is to swap the model distribution with the label distribution, i.e., the "empirical Fisher" matrix:

$$I_{\text{emp}}(w) = \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}}\left[g_{x,y}(w)g_{x,y}(w)^\top\right] = \mathop{\text{Cov}}_{(x,y)\sim\mathbf{pop}}\left(g_{x,y}(w)\right) + g(w)g(w)^\top \tag{59}$$

The approximation $I(w) \approx I_{\text{emp}(w)}$ is somewhat justified by the fact that $I(w) \to I_{\text{emp}}(w)$ assuming $w \to w^\star$, but in general it is highly flawed (see Section 2.1 of Grosse (2021)) and behaves strangely when used directly in (57). For instance, if $g_{x,y}(w)$ happens to have small covariance but large on average, we have $I_{\text{emp}}(w) \approx g(w)g(w)^\top$ which specifies an "inverse gradient scaling" $w \leftarrow w - \eta(g(w)g(w)^\top)^{-1}g(w)$ where the weights with the largest gradient values are updated the *least* (Kunstner *et al.*, 2019).[11] We can "fix" the inverse scaling by taking the square root,

$$I^{1/2}_{\text{emp}}(w) = \mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}}\left[g_{x,y}(w)g_{x,y}(w)^\top\right]^{1/2}$$

Further using the usual diagonal approximation $g_{x,y}(w)g_{x,y}(w)^\top \approx \text{diag}(g^2_{x,y}(w))$ for sparsity and using the linearity of diag, we have

$$I^{1/2}_{\text{emp,diag}}(w) = \text{diag}\left(\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}}\left[g^2_{x,y}(w)\right]\right)^{1/2} \tag{60}$$

Let $v \in \mathbb{R}^d$ denote a finite-sample estimator of $\mathbf{E}_{(x,y)\sim\mathbf{pop}}[g^2_{x,y}(w)]$. Using this estimator in $w \leftarrow w - \eta I^{-1/2}_{\text{emp,diag}}(w)g(w)$, we recover the per-parameter Adam update

$$w_j \leftarrow w_j - \frac{\eta}{\sqrt{v_j}}g(w_j)$$

Since $I^{1/2}_{\text{emp,diag}}(w)$ is viewable as an approximation of the Fisher $I(w)$ (not just the Hessian $H(w)$), Adam can be further motivated as an approximation of natural gradient descent $w \leftarrow w - I(w)^{-1}g(w)$ which optimizes the loss in an information-based transformation of the coordinate system (thereby invariant to the underlying geometry).

6.2 The Hessian View of Shampoo

Let $W \in \mathbb{R}^{m\times n}$ denote a weight matrix. Let $L : \mathbb{R}^{m\times n} \to \mathbb{R}$ be a random loss function for this weight (i.e., random in data). Let $G = \nabla L(W) \in \mathbb{R}^{m\times n}$. Let $w = \overline{\text{vec}}(W) \in \mathbb{R}^{mn}$ and $g = \overline{\text{vec}}(G) \in \mathbb{R}^{mn}$, with the corresponding reshaped loss $l : \mathbb{R}^{mn} \to \mathbb{R}$ defined as $l(\overline{\text{vec}}(W)) = L(W)$. Newton's method corresponds to $w \leftarrow w - \eta H^{-1}g$ where $H = \nabla^2 l(w) \in \mathbb{R}^{mn\times mn}$. Again we approximate $H \approx I_{\text{fisher}} \approx I^{1/2}_{\text{emp}}$ where $I_{\text{emp}} = \mathbf{E}[gg^\top] \in \mathbb{R}^{mn\times mn}$. Inverting and multiplying by an $mn \times mn$ matrix incurs a $O(m^3n^3)$ runtime overhead. We can reduce this to $O(m^3 + n^3)$ by considering a *Kronecker decomposition*:

$$A_\star, B_\star = \mathop{\arg\min}_{A\in\mathbb{R}^{m\times m},\ B\in\mathbb{R}^{n\times n}} ||I_{\text{emp}} - A \otimes B||_F \tag{61}$$

since the corresponding update $w \leftarrow w - \eta(A_\star^{-1/2} \otimes B_\star^{-1/2})g$ can be achieved by double-sided matrix multiplication (see (99))

$$W \leftarrow W - \eta A_\star^{-1/2} G (B_\star^{-1/2})^\top \tag{62}$$

We can solve (61) by a rank-1 SVD of a $mn \times mn \mapsto m^2 \times n^2$ rearrangement of the target matrix I_{emp}. Specifically, let $\widetilde{I}_{\text{emp}} = \textbf{rearrange}(I_{\text{emp}}) \in \mathbb{R}^{m^2\times n^2}$ where the function **rearrange** is defined in (102). Let $u_1 \in \mathbb{R}^{m^2}$ and

[11] The problem worsens if we use the batched gradient estimator $g_B(w) = \frac{1}{|B|}\sum_{(x,y)\in B}\frac{\partial L(f_w(x),y)}{\partial w}$ in (59) (which is closer to the practice). Since $\mathbf{E}[g_B(w)] = g(w)$ and $\text{Cov}_B(g_B(w)) = \frac{1}{|B|}\text{Cov}_{x,y}(g_{x,y}(w))$, the empirical Fisher estimator using the batched gradient estimator becomes

$$\mathop{\mathbf{E}}_{B\sim\mathbf{pop}^{|B|}}\left[g_B(w)g_B(w)^\top\right] = \frac{1}{|B|}I_{\text{emp}}(w) + \left(1 - \frac{1}{|B|}\right)g(w)g(w)^\top$$

which shows that $g(w)g(w)^\top$ dominates the estimate as the batch size grows.

$v_1 \in \mathbb{R}^{n^2}$ be the top left and right singular vectors of $\widetilde{I}_{\text{emp}}$. Let $U_1 \in \mathbb{R}^{m\times m}$ and $V_1 \in \mathbb{R}^{n\times n}$ denote the row-major matrix representations of u_1, v_1. Then for some $a, b > 0$ (Corollary G.2):

$$A_\star = aU_1 \qquad\qquad B_\star = bV_1$$

We can easily verify that $\widetilde{I}_{\text{emp}} = \mathbf{E}[G \otimes G]$ using (101) and (103). We can estimate the top singular vectors of $\widetilde{I}_{\text{emp}}$ by the power method: use some initial $l_0 \in \mathbb{R}^{m^2}$ and $r_0 \in \mathbb{R}^{n^2}$ and repeat $l_{i+1} = \widetilde{I}_{\text{emp}} r_i$ and $r_{i+1} = (\widetilde{I}_{\text{emp}})^\top l_i$. It is well known that $\frac{l_i}{||l_i||} \to u_1$ and $\frac{r_i}{||r_i||} \to v_1$ (assuming $\sigma_1 > \sigma_2$ for simplicity). Now choose $l_0 = \overline{\text{vec}}(I_m)$ and $r_0 = \overline{\text{vec}}(I_n)$. Then one iteration yields

$$l_1 = \mathbf{E}[G \otimes G] r_0 = \overline{\text{vec}}(\mathbf{E}[GG^\top])$$
$$r_1 = \mathbf{E}[G \otimes G]^\top l_0 = \overline{\text{vec}}(\mathbf{E}[G^\top G])$$

Treating these as rough estimates of (some scaling of) u_1 and v_1, we can argue that using (61) in (62) yields (with an appriate η)

$$W \leftarrow W - \eta \mathbf{E}[GG^\top]^{-1/2}\, G\, \mathbf{E}[G^\top G]^{-1/2} \tag{63}$$

This corresponds to using the square of the Shampoo preconditioner $A^2_{\text{shampoo}} = L^{1/2} \otimes R^{1/2}$, since Shampoo with EMA specifies (Section 5.1)

$$W \leftarrow W - \eta \mathbf{E}[GG^\top]^{-1/4}\, G\, \mathbf{E}[G^\top G]^{-1/4}$$

Morwani *et al.* (2024) justify the identity initialization as a way of making $\cos(l_1, u_1) = \frac{(v_1^\top r_0)\sigma_1}{\sqrt{(v_i^\top r_0)^2 \sigma_i^2}}$ closer to 1 (similarly for r_1, v_1). Specifically, they show that $r_0 = \overline{\text{vec}}(I_n)$ yields $v_1^\top r_0 > v_i^\top r_0$ for $i \geq 2$.

6.2.1 Exact decomposition

Lemma 6.1 (Morwani *et al.* (2024))**.** If $\widetilde{I}_{\text{emp}} = \textbf{rearrange}(I_{\text{emp}}) = \mathbf{E}[G \otimes G] \in \mathbb{R}^{m^2 \times n^2}$ is rank-1,

$$I_{\text{emp}} = \frac{\mathbf{E}[GG^\top] \otimes \mathbf{E}[G^\top G]}{\text{tr}\left(\mathbf{E}[GG^\top]\right)} \in \mathbb{R}^{mn \times mn} \tag{64}$$

Proof. Let $\widetilde{I}_{\text{emp}} = \sigma u v^\top$ be a rank-1 SVD. This implies $I_{\text{emp}} = \sigma U \otimes V$ where $u = \overline{\text{vec}}(U)$ and $v = \overline{\text{vec}}(V)$. Shampoo's iteration gives us $\widetilde{I}_{\text{emp}} r_0 = \sigma u v^\top r_0 = \overline{\text{vec}}(\mathbf{E}[GG^\top])$ where $r_0 = \overline{\text{vec}}(I_n)$. Let $v_\perp = r_0 - \text{Proj}_{\text{span}(v)}(r_0)$ where

$$\text{Proj}_{\text{span}(v)}(r_0) = vv^\top r_0 = (\overline{\text{vec}}(I_n)^\top \overline{\text{vec}}(V))v = \text{tr}\,(V)\, v$$

Then

$$\overline{\text{vec}}(\mathbf{E}[GG^\top]) = \sigma u v^\top (v_\perp + \text{tr}\,(V)\, v) = \sigma \text{tr}\,(V)\, u$$
$$\mathbf{E}[GG^\top] = \sigma \text{tr}\,(V)\, U$$

which also implies $\text{tr}\left(\mathbf{E}[GG^\top]\right) = \sigma \text{tr}\,(V) \,\text{tr}\,(U)$. Similarly, $\mathbf{E}[G^\top G] = \sigma \text{tr}\,(U)\, V$. We now obtain (64) by re-expressing $I_{\text{emp}} = \sigma U \otimes V$. □

The rank-1 assumption in Lemma 6.1 is unrealistically strong (likely holds only for linear logistic regressor where $G \in \mathbb{R}^{m\times 1}$). But it suggests the following "idealized" shampoo iteration which corresponds to the Newton step $w \leftarrow w - \eta H^{-1} g$ on $w = \overline{\text{vec}}(W) \in \mathbb{R}^{mn}$ using $H \approx I_{\text{fisher}} \approx I_{\text{emp}}^{1/2} = \frac{1}{\sqrt{\text{tr}(\mathbf{E}[GG^\top])}} \mathbf{E}[GG^\top]^{1/2} \otimes \mathbf{E}[G^\top G]^{1/2}$.

IdealizedShampooIteration
Input: Current $W \in \mathbb{R}^{m\times n}$, random loss $L : \mathbb{R}^{m\times n} \to \mathbb{R}$, learning rate $\eta > 0$

1. Compute $G = \nabla L(W) \in \mathbb{R}^{m\times n}$.
2. Update the estimates $\mathbf{E}[GG^\top] \in \mathbb{R}^{m\times m}$ and $\mathbf{E}[G^\top G] \in \mathbb{R}^{n\times n}$ (e.g., bias-corrected EMA).
3. $W \leftarrow W - \eta \sqrt{\text{tr}\,(\mathbf{E}[GG^\top])}\, \mathbf{E}[GG^\top]^{-1/2}\, G\, \mathbf{E}[G^\top G]^{-1/2}$

7 Muon

Muon falls out of first principles. Let $J : \mathbb{R}^{m\times n} \to \mathbb{R}$ denote the loss function. At a current point $Z \in \mathbb{R}^{m\times n}$ with gradient $G = \nabla J(Z) \in \mathbb{R}^{m\times n}$, the usual ways to get the next point by local search are

$$X_{\text{TR}} = \underset{X\in\mathbb{R}^{m\times n}:\ ||X-Z||\le\eta}{\arg\min} \quad J(Z) + \langle G, X - Z\rangle \tag{65}$$

$$X_{\text{MM}} = \underset{X\in\mathbb{R}^{m\times n}}{\arg\min} \quad J(Z) + \langle G, X - Z\rangle + \frac{1}{2\eta}\,||X - Z||^2 \tag{66}$$

for some choice of norm $||\cdot|| : \mathbb{R}^{m\times n} \to \mathbb{R}$ and "step size" $\eta > 0$. (65) is the trust-region principle (minimize a local approximation within a ball of radius η). (66) is the majorization-minimization principle (minimize a surrogate upper bound, specifically the quadratic-penalty form assuming smoothness—see Appendix P). Expressed in the update $\Delta = X - Z$, (65) and (66) become

$$\Delta_{\text{TR}} = \underset{\Delta\in\mathbb{R}^{m\times n}:||\Delta||\le\eta}{\arg\min} \quad \langle G, \Delta\rangle \tag{67}$$

$$\Delta_{\text{MM}} = \underset{\Delta\in\mathbb{R}^{m\times n}}{\arg\min} \quad \langle G, \Delta\rangle + \frac{1}{2\eta}\,||\Delta||^2 \tag{68}$$

Committing to specific norm yields closed-form updates.[12]

$$\Delta_{\text{TR}} = \begin{cases} -\frac{\eta}{||G||_F}G & \text{if } ||\Delta|| = ||\Delta||_F \\ -\eta\,\textbf{sign}\,(G) & \text{if } ||\Delta|| = ||\Delta||_{\infty,\text{elt}} \\ -\eta UV^\top & \text{if } ||\Delta|| = ||\Delta||_2 \end{cases} \qquad \Delta_{\text{MM}} = \begin{cases} -\eta G & \text{if } ||\Delta|| = ||\Delta||_F \\ -\eta\,||G||_{1,\text{elt}}\,\textbf{sign}\,(G) & \text{if } ||\Delta|| = ||\Delta||_{\infty,\text{elt}} \\ -\eta\,||G||_{\text{nuc}}\,UV^\top & \text{if } ||\Delta|| = ||\Delta||_2 \end{cases} \tag{69}$$

where $G = U\Sigma V^\top$ is the SVD of G. For the same norm, TR and MM yield the same direction. If $G \in \mathbb{R}^{m\times 1}$ is a vector, we recover gradient descent under $||\cdot||_F = ||\cdot||_2$ and sign descent under $||\cdot||_\infty$ (focusing on the direction only). However, for $n > 1$ the spectral norm yields the **Muon direction** $UV^\top \in \mathbb{R}^{m\times n}$.

Proof of (69). (TR): The case for $||\cdot||_F$ and $||\cdot||_{\infty,\text{elt}}$ can be read off of (67) for $||\cdot||_F$ and $||\cdot||_{\infty,\text{elt}}$. The case for $||\cdot||_2$ follows immediately from von Neumann's trace inequality $\langle G, \Delta\rangle \ge -\eta\,||G||_{\text{nuc}}$ (for a feasible Δ) and the fact that $\Delta = -\eta UV^\top$ makes it tight. (MM): Since $||\cdot||$ may not be differentiable, we move it out of the objective by reparameterizing $\Delta = rO$ where $r \ge 0$ is the size and O is the direction with $||O|| = 1$ and rewrite (68) as

$$r^\star, O^\star = \underset{r\ge 0,\ O\in\mathbb{R}^{m\times n}:\ ||O||=1}{\arg\min} \quad r\,\langle G, O\rangle + \frac{r^2}{2\eta}$$

For fixed r, the minimizing direction is any $O^\star$ such that $\langle G, O^\star\rangle = -\,||G||_*$ (i.e., dual norm definition). Thus the minimizing size always satisfies $r^\star = \arg\min_{r\ge 0} -r\,||G||_* + \frac{r^2}{2\eta} = \eta\,||G||_*$. Now the updates for Δ_{MM} follow from the dual norm relations $||\cdot||_F \leftrightarrow ||\cdot||_F$, $||\cdot||_{1,\text{elt}} \leftrightarrow ||\cdot||_{\infty,\text{elt}}$, and $||\cdot||_{\text{nuc}} \leftrightarrow ||\cdot||_2$ with the observations already made for (TR). □

Importantly, computing the Muon direction at each step is practically feasible thanks to specialized algorithms such as Newton-Schulz iteration (Appendix Q.3).

8 Manifold Optimization

Manifold optimization (MO) explicitly maintains each weight unit on a nontrivial manifold, typically the Euclidean sphere if vector-valued and the Stiefel manifold if matrix-valued.[13] It can be viewed as "hard" regularization which may impact generalization in beneficial ways.

[12] For infinity- and 1-norm of a matrix, we use the elementwise version for "backward compatibility" with vector inputs. We can also express the former as an induced operator norm $||A||_{1\to\infty} = \max_{x:||x||_1\le 1} ||Ax||_\infty = \max_{i,j} |A_{i,j}| = ||A||_{\infty,\text{elt}}$.

[13] A manifold is a topological space that *locally* resembles Euclidean space. Thus $\mathbb{R}^d$ is trivially a manifold; a hypersphere $\{u \in \mathbb{R}^d : ||u||_2 = 1\}$ is a manifold ("locally flat, globally curved"). With matrices in $\mathbb{R}^{D\times d}$ ($D \ge d$), the Stiefel manifold $\{A \in \mathbb{R}^{D\times d} : A^\top A = I_{d\times d}\}$ is a standard nontrivial manifold.

8.1 Vector Case

Let $w \in \mathbb{R}^d$ denote a weight vector (e.g., a row of layer $W \in \mathbb{R}^{D\times d}$ where $Wx \in \mathbb{R}^D$ is the output). At every training step t, we keep $||w_t||_2 = r$ for some radius $r > 0$ by "retracting each update back to the hypersphere" (i.e., renormalize length).

8.1.1 Pin-to-sphere

The simplest approach is vanilla projected gradient step, so-called "pin-to-sphere". Take the usual gradient step $w_t - \eta_t g_{w_t}$ with learning rate $\eta_t > 0$ where $g_{w_t} = \partial \text{Loss}_t / \partial w_t \in \mathbb{R}^d$ is the "ambient" gradient, and retract:

$$w_{t+1} = \frac{r}{||w_t - \eta_t g_{w_t}||_2}(w_t - \eta_t g_{w_t}) \tag{70}$$

A potential downside of this formulation is that g_{w_t} may have a large component parallel to w_t, which is annihilated by retraction. In that case, it is visually clear that even if we take a huge step in the ambient space we end up with little movement on the manifold, making it difficult to control the effective update size with the learning rate.

Preconditioning view. Pin-to-sphere locally looks like preconditioned gradient descent using the projection matrix onto span $(w_t)^\perp$ as the (per-step and singular) preconditioner (Lemma S.19):

$$w_{t+1} = w_t - \eta_t g^\perp_{w_t} + O(\eta_t^2) \qquad\qquad g^\perp_{w_t} := \left(I_d - \frac{1}{r^2} w_t w_t^\top\right) g_{w_t} \tag{71}$$

Intuitively, we learn from only what is directionally new.

8.1.2 Classical formulation

We can just take a gradient step directly in the tangent space. This gives the classical MO update

$$w_{t+1} = \frac{r}{\sqrt{r^2 + \eta_t^2 \left|\left|g^\perp_{w_t}\right|\right|_2^2}}(w_t - \eta_t g^\perp_{w_t}) \tag{72}$$

where the normalization uses the fact that $w_t^\top g^\perp_{w_t} = 0$ and $||w_t||_2 = r$. The same first-order preconditioning view (71) remains valid. While this makes the learning rate correspond to the movement in the tangent space rather than ambient, it still depends on $g^\perp_{w_t}$ which may be tiny.

8.1.3 Trust-region formulation

We may explicitly match the tangent step size with the learning rate. That is, first find a step satisfying

$$\Delta_t = \underset{\Delta\in\mathbb{R}^d:\ ||\Delta||\le\eta_t,\ \Delta^\top w_t = 0}{\arg\max} \Delta^\top g_{w_t} \tag{73}$$

then retract $w_t - \Delta_t$. (73) is now "just" steepest descent (aka. trust-region method) with an extra orthogonality constraint, so we can derive various updates by changing the norm $||\cdot||$. With the l_2 norm, the solution is $\Delta_t = \eta_t g^\perp_{w_t} / \left|\left|g^\perp_{w_t}\right|\right|_2$ (i.e., rescale the classical MO). For general norm, we can solve (73) by Lagrangian duality:

$$\lambda_t^\star \in \underset{\lambda\in\mathbb{R}}{\arg\min} ||g_{w_t} + \lambda w_t||_* \qquad\qquad \Delta_t = \eta_t z_t$$

where $z_t \in \mathbb{R}^d$ is a subgradient $z_t \in \partial ||g_{w_t} + \lambda_t^\star w_t||_*$ orthogonal to w_t, which must exist (Lemma S.21). We can easily check that this recovers the l_2 solution as a special case.

8.1.4 Weight normalization

Weight normalization (WN) (Salimans and Kingma, 2016) reparameterizes the weight $w_t \in \mathbb{R}^d$ as "direction" $v_t \in \mathbb{R}^d$ and "scale" $s_t \in \mathbb{R}$:

$$w_t = s_t \frac{v_t}{||v_t||_2}$$

The intuition is that while model expressiveness does not change, (v_t, s_t) get more fine-grained gradients. Given the gradient $g_{w_t} \in \mathbb{R}^d$, by the chain rule

$$g_{v_t} = \frac{s_t}{||v_t||_2}\left(I_d - \bar{v}_t\bar{v}_t^\top\right) g_{w_t} = \frac{s_t}{||v_t||_2} g_{w_t}^\perp \qquad\qquad g_{s_t} = \bar{v}_t^\top g_{w_t} \tag{74}$$

which uses the well-known Jacobian $J_F(v) = (1/||v||_2)(I_d - \bar{v}\bar{v}^\top)$ of $F(v) = \bar{v} := v/||v||_2$ and $\bar{v}_t = \bar{w}_t$. Note that $g_{v_t}^\top v_t = 0$ (i.e., v_t indeed only learns direction). For gradient descent, WN corresponds to preconditioning the update for w_t. To see why, the updated w_t takes the form

$$w_{t+1} = (s_t + \Delta s_t)(\bar{v}_t + \Delta\bar{v}_t)$$

Taking the first-order approximation $\Delta w_t \approx s_t\Delta\bar{v}_t + \Delta s_t\bar{v}_t$ and plugging $\Delta s_t = -\eta_t g_{s_t}$ and $\Delta\bar{v}_t \approx J_F(v_t)\Delta v_t = -\eta_t J_F(v_t) g_{v_t}$[14] with (74), we have

$$\Delta w_t \approx -\eta_t\left(\frac{s_t^2}{||v_t||_2^2}\left(I_d - \bar{v}_t\bar{v}_t^\top\right) + \bar{v}_t\bar{v}_t^\top\right) g_{w_t} \tag{75}$$

So the update for w_t is guided by v_t, either amplifying or shrinking the component of g_{w_t} lying in span $(v_t)^\perp$ (we recover vanilla gradient descent, whereas if $s_t = ||v_t||_2$). It is worth noting that if we freeze $s_t = r$, $\Delta w_t \propto -g_{w_t}^\perp$ resembles the classical MO update (72). The precise relationship is given below.

Lemma 8.1. Let $w_{t+1} = r\bar{v}_{t+1}$ denote the materialized weight under WN using frozen $s_t = r$. Then

$$w_{t+1} = \frac{r}{\sqrt{r^2 + \alpha_t^2\left|\left|g_{w_t}^\perp\right|\right|_2^2}}(w_t - \alpha_t g_{w_t}^\perp) \qquad\qquad \alpha_t = \eta_t\frac{r^2}{||v_t||_2^2} \tag{76}$$

We leave the proof as an exercise.[15] Comparing (72) with (76), we conclude that frozen-scale WN is equivalent to classical MO with an adaptive learning rate (controlled by $||v_t||_2$). A useful mental summary is

- MO locally assigns tangential factor 1 and radial factor 0 (71).
- WN locally assigns tangential factor $\frac{s_t^2}{||v_t||_2^2}$ and radial factor 1 (75).
- Thus with frozen scale and after accounting for the effective learning rate, WN collapses to MO (76).

8.2 Matrix Case

Let $W \in \mathbb{R}^{D\times d}$ denote a layer weight with $D \geq d$. We keep $W_t \in \left\{A \in \mathbb{R}^{D\times d} : A^\top A = I_{d\times d}\right\}$ by retracting each update $W_t - \Delta_t$ back to the manifold. A matrix-analog of the Euclidean projection is

$$W_{t+1} = \underset{W\in\mathbb{R}^{D\times d}:\, W^\top W = I_{d\times d}}{\arg\min} ||W - (W_t - \Delta_t)||_F = \mathrm{Orth}(W_t - \Delta_t) \tag{77}$$

where $\mathrm{Orth}(M) = UV^\top$ computes the orthogonalization of $M = U\Sigma V^\top$ which is numerically convenient (e.g., Newton-Schulz iteration Q.3.1). The trust-region formulation (73) gives us: given the gradient $G_t \in \mathbb{R}^{D\times d}$, find

$$\Delta_t = \underset{\Delta\in\mathbb{R}^{D\times d}:\, ||\Delta||=\eta_t,\, \Delta^\top W_t + W_t^\top\Delta = 0_{d\times d}}{\arg\max} \langle\Delta, G_t\rangle \tag{78}$$

which uses the known characterization of the tangent space of the Stiefel manifold (Tagare, 2011). We can again solve (78) by Lagrangian duality:

$$\Lambda_t^\star \in \underset{\Lambda\in\mathrm{Sym}(d)}{\arg\min} ||G_t + W_t\Lambda||_* \qquad\qquad \Delta_t = \eta_t Z_t \tag{79}$$

where $Z_t \in \partial\,||G_t + W_t\Lambda_t^\star||_*$ satisfies $Z_t^\top W_t + W_t^\top Z_t = 0_{d\times d}$ and must exist (Lemma S.22). Bernstein (2025) chooses the spectral norm $||\cdot||_2$ in (78)[16] and approximates (79) by subgradient descent on the convex objective $\left|\left|G_t + W_t(\Lambda + \Lambda^\top)\right|\right|_{\mathrm{nuc}}$ over $\Lambda \in \mathbb{R}^{d\times d}$ with a stopping criterion to promote tangency.

[14] This uses the first-order approximation $\bar{v} + \Delta\bar{v} = F(v + \Delta v) \approx \bar{v} + J_F(v)\Delta v$.

[15] Use (74) to express $v_{t+1} = v_t - \eta_t g_{v_t} = (||v_t||_2/r)(w_t - \alpha_t g_{w_t}^\perp)$. Normalizing by $||v_{t+1}||_2$ yields (76).

[16] Referred to as "Manifold Muon" since it becomes the Muon update $\Delta_t = -\eta_t U_t V_t^\top$ (67) if we omit the tangent-space constraint.

Pointers

- Introduction to Online Learning by Francesco Orabona, in particular online gradient descent and adaptive algorithms
- Lecture slides by Sham Kakade
- Lecture slides by Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky
- Blog by Sebastian Ruder
- Notes on mirror descent by Xinhua Zhang
- Course notes by Roger Grosse

References

Bernstein, J. (2025). Modular manifolds. *Thinking Machines Lab: Connectionism*. https://thinkingmachines.ai/blog/modular-manifolds/.

Botev, A., Lever, G., and Barber, D. (2017). Nesterov's accelerated gradient and momentum as approximations to regularised update descent. In *2017 International joint conference on neural networks (IJCNN)*, pages 1899–1903. IEEE.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, **12**(7).

Everett, K. E., Xiao, L., Wortsman, M., Alemi, A. A., Novak, R., Liu, P. J., Gur, I., Sohl-Dickstein, J., Kaelbling, L. P., Lee, J., and Pennington, J. (2024). Scaling exponents across parameterizations and optimizers. In *Forty-first International Conference on Machine Learning*.

Finesso, L. and Spreij, P. (2006). Nonnegative matrix factorization and i-divergence alternating minimization. *Linear Algebra and its Applications*, **416**(2-3), 270–287.

Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, **22**(4), 1469–1492.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Grosse, R. (2021). Adaptive gradient methods, normalization, and weight decay. https://www.cs.toronto.edu/~rgrosse/courses/csc2541_2021/readings/L05_normalization.pdf.

Gupta, V., Koren, T., and Singer, Y. (2018). Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR.

Hanson, S. and Pratt, L. (1988). Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, **1**.

Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., and Bernstein, J. (2024). Muon: An optimizer for hidden layers in neural networks.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kunstner, F., Hennig, P., and Balles, L. (2019). Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, **32**.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, **401**(6755), 788–791.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Morwani, D., Shapira, I., Vyas, N., Malach, E., Kakade, S., and Janson, L. (2024). A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*.

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. In *Dokl. Akad. Nauk. SSSR*, volume 269, page 543.

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, **29**.

Sankar, A. R., Khasbage, Y., Vigneswaran, R., and Balasubramanian, V. N. (2021). A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9481–9488.

Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Shi, H.-J. M., Lee, T.-H., Iwasaki, S., Gallego-Posada, J., Li, Z., Rangadurai, K., Mudigere, D., and Rabbat, M. (2023). A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale. *arXiv preprint arXiv:2309.06497*.

Tagare, H. D. (2011). Notes on optimization on stiefel manifolds. *Yale University, New Haven*.

Tieleman, T., Hinton, G., *et al.* (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, **4**(2), 26–31.

Van Loan, C. F. and Pitsianis, N. (1993). *Approximation with Kronecker products*. Springer.

Wortsman, M., Liu, P. J., Xiao, L., Everett, K. E., Alemi, A. A., Adlam, B., Co-Reyes, J. D., Gur, I., Kumar, A., Novak, R., Pennington, J., Sohl-Dickstein, J., Xu, K., Lee, J., Gilmer, J., and Kornblith, S. (2024). Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*.

Zhuang, Z., Liu, M., Cutkosky, A., and Orabona, F. (2022). Understanding adamw through proximal methods and scale-freeness. *arXiv preprint arXiv:2202.00089*.

A Convex Stochastic Optimization (CSO)

Let $x \sim \mathbf{pop}$ define a convex per-example loss $J_x(w) \in \mathbb{R}$ over $w \in V$. Let

$$w^\star = \arg\min_{w \in V} \mathop{\mathbf{E}}_{x \sim \mathbf{pop}} [J_x(w)] \tag{80}$$

Denote the expected loss $J(w) = \mathbf{E}_{x \sim \mathbf{pop}}[J_x(w)]$ (which remains convex) and let $J^\star = J(w^\star)$. The goal of first-order CSO may be stated as developing a gradient-based algorithm that produces a proposal $w \in V$ given T iid samples from $\mathbf{pop}$ (note that w may be itself random) such that

$$E_w[J(w)] - J^\star \leq O\left(\frac{1}{T^\alpha}\right) \tag{81}$$

for some $\alpha > 0$ (the higher the faster convergence). We may use OCO to solve CSO. Starting from some initial $w_1 \in V$, for $t = 1 \dots T$, we sample an iid $x_t \sim \mathbf{pop}$, get punished by the convex loss $J_{x_t}(w_t) \in \mathbb{R}$, and obtain w_{t+1} from the algorithm. This makes $w_t \in V$ random in the past data $x_0 = \varnothing, x_1 \dots x_{t-1}$. Let $B(T)$ be a regret bound of the OCO algorithm. Then

$$\sum_{t=1}^T J_{x_t}(w_t) - \min_{w \in V} \sum_{t=1}^T J_{x_t}(w) \leq B(T) \tag{82}$$

In particular,

$$\sum_{t=1}^T J_{x_t}(w_t) - \sum_{t=1}^T J_{x_t}(w^\star) \leq B(T)$$

Dividing both sides by T yields

$$\frac{1}{T}\sum_{t=1}^T J_{x_t}(w_t) - \frac{1}{T}\sum_{t=1}^T J_{x_t}(w^\star) \leq \frac{B(T)}{T}$$

Take the expectation wrt. the iid samples $x_1 \dots x_T \sim \mathbf{pop}$ on both sides. The second term on the LHS becomes $J^\star$. The first term is

$$\begin{aligned}
\mathop{\mathbf{E}}_{x_1 \dots x_T \sim \mathbf{pop}} \left[\frac{1}{T}\sum_{t=1}^T J_{x_t}(w_t)\right] &= \frac{1}{T}\sum_{t=1}^T \mathop{\mathbf{E}}_{x_1 \dots x_t \sim \mathbf{pop}} [J_{x_t}(w_t)] \\
&= \frac{1}{T}\sum_{t=1}^T \mathop{\mathbf{E}}_{w_t} [J(w_t)] && \text{(since } w_t \text{ is independent of } x_t\text{)} \\
&= \mathop{\mathbf{E}}_{w_1 \dots w_T} \left[\frac{1}{T}\sum_{t=1}^T J(w_t)\right] \\
&\geq \mathop{\mathbf{E}}_{w_1 \dots w_T} \left[J\left(\frac{1}{T}\sum_{t=1}^T w_t\right)\right] && \text{(convexity of } J\text{)}
\end{aligned}$$

Thus we have

$$\mathop{\mathbf{E}}_{w_1 \dots w_T} \left[J\left(\frac{1}{T}\sum_{t=1}^T w_t\right)\right] - J^\star \leq \frac{B(T)}{T}$$

I.e., taking the average of $w_1 \dots w_T$ from the online algorithm on (any sequence of) iid samples $x_1 \dots x_T \sim \mathbf{pop}$ yields a solution that on average (wrt. sampling randomness) falls behind $J^\star$ at the rate of $O(\frac{B(T)}{T})$. In particular, if $B(T) = o(T)$ (i.e., sublinear regret), the solution is guaranteed to converge to $J^\star$ asymptotically. For instance, if $B(T) = O(\sqrt{T})$, the solution has the convergence rate of $O(\frac{1}{\sqrt{T}})$.

A.1 Optimal First-Order Convergence Rate

While OCO can be used to achieve (81) up to $\alpha = \frac{1}{2}$ (limited by its regret lower bound $\Omega(\sqrt{T})$), it does not exploit the fact that the loss is not adversarial (J_x is determined randomly by $x \sim \mathbf{pop}$). We may achieve $\alpha = 1$ (using SGD) if J is L-smooth and the gradients are noiseless (i.e., infinite batch size). Specifically, if we perform the true gradient steps $w_{t+1} = w_t - (1/L)\nabla J(w_t)$ (i.e., this is GD, not SGD), we have (Lemma S.5)

$$J(w_{T+1}) - J^\star \leq \frac{L\,||w_1 - w^\star||_2^2}{2T}$$

How far can we push α using only the gradient information (i.e., first-order)? Nesterov (1983) propose a method that achieves $\alpha = 2$. Again assuming J is L-smooth and the gradients are noiseless, we update

$$w_{t+1} = y_t - (1/L)\nabla J(y_t)$$
$$y_{t+1} = w_{t+1} + \frac{t}{t+3}(w_{t+1} - w_t)$$

which is equivalent to heavy-ball gradient descent with a lookahead gradient (Section 2.3). Then it holds that

$$J(w_T) - J^\star \leq \frac{2L\,||w_1 - w^\star||_2^2}{(T+1)^2}$$

Nemirovskij and Yudin (1983) show that $O(1/T^2)$ is a lower bound for every first-order method, hence $\alpha = 2$ is the best possible in CSO when one only uses gradients.

Noisy gradients. If the (unbiased) gradients are noisy $g_t = \nabla J_{x_t}(w_t)$ with $x_t \sim \mathbf{pop}$ but in a bounded way $\mathbf{E}_{x_t}[||g_t - \nabla J(w_t)||^2] \leq \sigma^2$, we can show that

$$\mathbf{E}[J(\bar{w}_T)] - J^\star \leq \underbrace{O\left(\frac{1}{T^2}\right)}_{\text{bias}} + \underbrace{O\left(\frac{\sigma}{\sqrt{T}}\right)}_{\text{variance}}$$

where $\bar{w}_T$ is some time-weighted average (Ghadimi and Lan, 2012). (The bias term is $O(1/T)$ for SGD.) Hence the $O(\sigma/\sqrt{T})$ variance term dominates the overall rate unless the gradients are noise-free.

A.2 Second-Order Convergence Rate

If we use the second-order information (i.e., the Hessian), we can actually achieve a fundamentally faster rate than (81). The following theorem describes the convergence behavior of Newton around the optimum for an idealized function (Corollary S.7, S.9; Lemma S.8):

Theorem A.1 (Convergence rate of Newton)**.** Assume $J : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex and M-smooth. Let $w^\star = \arg\min_{w \in \mathbb{R}^d} J(w)$ denote the unique minimizer. Define the **Newton ball** as $\mathcal{B} = \left\{w : ||w - w^\star||_2 \leq \frac{m}{L}\right\}$. Assume that the Hessian is L-Lipschitz in the Newton ball in operator norm for $||\cdot||_2$ (i.e., $\left|\left|\nabla^2 J(u) - \nabla^2 J(v)\right|\right|_2 \leq L\,||u - v||_2$ for $u, v \in \mathcal{B}$). If $w_t \in \mathcal{B}$, and we perform the Newton steps $w_{t+1} = w_t - (\nabla^2 J(w_t))^{-1}\nabla J(w_t)$:

1. (Linear phase) For all $k \geq 1$, we have

$$J(w_{t+k}) - J^\star \leq \left(\frac{Mm^2}{2L^2}\right) 4^{-k}$$

   with $||w_{t+k} - w^\star||_2 \leq \left(\frac{m}{L}\right) 2^{-k}$.

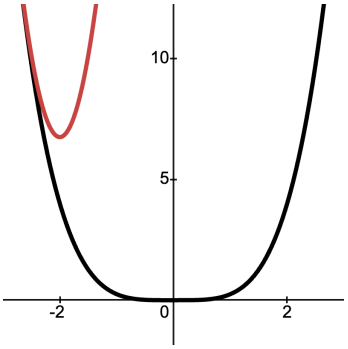
2. (Quadratic phase) Let $t' = t + \left\lceil \log_4\left(\frac{M^2}{2m^2}\right)\right\rceil$. For all $k \geq 1$, we have

$$J(w_{t'+k}) - J^\star \leq \left(\frac{2m^4}{ML^2}\right) 2^{-2^k}$$

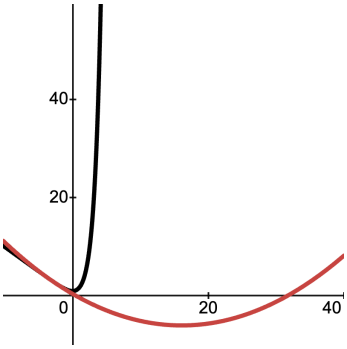
So around the optimum with a well-behaved Hessian, Newton achieves an exponential rate $O(4^{-T})$ and after some burn-in a *doubly*-exponential rate $O(2^{-2^T})$, which decays far faster than the first-order rate $O(T^{-\alpha})$ where $\alpha = 1$ for SGD and $\alpha = 2$ for Nesterov.

A.2.1 Failure mode of Newton

Theorem A.1 crucially depends on the bounded movement of the Hessian (i.e., Lipschitz). Outside the Newton ball, if the curvature changes too much, Newton may lose its quadratic advantage. A simple 1D convex example is $J(w) = \frac{1}{4}w^4$ with $w^\star = 0$ and $J^\star = 0$. Note the Hessian $J''(w) = 3w^2$ may not be Lipschitz far away from 0 (i.e., $3\left|u^2 - v^2\right| \gg |u - v|$). Newton says $w_{t+1} = \frac{2}{3}w_t$, yielding $w_{T+1} = (\frac{2}{3})^T w_1$ and $J(w_{T+1}) - J^\star = O((81/16)^{-T})$, never entering the quadratic phase $O(2^{-2^T})$ (still faster than SGD's $O(T^{-1})$). Worse, Newton may diverge. Consider $J(w) = e^w - w$ with $w^\star = 0$ and $J^\star = 1$. Newton says $w_{t+1} = w_t - 1 + e^{-w_t}$. If $w_t = -100$, then $w_{t+1} > 10^{43}$ and the subsequent steps will not be able to bring it back to 0 (e.g., $w_{t+2} \approx 10^{43} - 1$), essentially failing to converge. The second-order approximations (which Newton minimizes) for these examples are visualized below at $w = -3$:



$J(w) = \frac{1}{4}w^4$



$J(w) = e^w - w$

Thus classical Newton necessiates "globalization" techniques to handle locations outside the Newton ball (e.g., do first-order until close to convergence, damp with a learning rate $\eta_t < 1$). When we view adaptive methods like AdaGrad, Adam, and Shampoo as approximating the inverse Hessian, we should note that they naturally simulate the damping scheme by the virtue of slowly moving preconditioners (typically some momentum or moving average) as well as explicit learning rate schedules.

A.2.2 Clarification on the rate terminology

We have shown the following convergence rates for non-stochastic (i.e., using the exact gradient/Hessian) convex optimization with well-behaved loss functions:

Method	$J(w_T) - J^\star$	Rate terminology
Gradient descent	$O(\frac{1}{T})$	Sublinear
Nesterov	$O(\frac{1}{T^2})$	Sublinear
Newton during linear phase	$O(\beta^T)$ where $\beta < 1$	Linear
Newton during quadratic phase	$O(\rho^{2^T})$ where $\rho < 1$	Quadratic (super-linear)

The adjectives are from local Q-convergence theory. Denoting the error at step t as E_t,

- $E_{t+1} = \beta E_t$ for $\beta < 1$ is called **linear**. This implies the exponential decay $O(\beta^T)$.
- $E_{t+1} = CE_t^2$ for $C > 0$ is called **quadratic**. This (asymptotically) implies the doubly exponential decay $O(\rho^{2^T})$ for some $\rho < 1$ (e.g., 2^{-2^T} after a burn-in).
- Any slower rate is called **sublinear**. This includes the polynomial decay $O(T^{-\alpha})$.

This is unrelated to the term sublinear regret in online learning, where the regret bound $o(T)$ is literally sublinear.

B Lower Bound on Regret

Let $V = \{\pm 1\}$ denote the hypothesis space. At each step t, the enemy *randomly* picks $x_t = \pm 1$ and defines the (linear) loss $l_t(w_t) = -x_t w_t$. Then no matter what $w_1 \dots w_T$ we choose, our expected cumulative loss is always zero by the linearity of expectation and the independence of w_t and x_t

$$\mathop{\mathbf{E}}_{x_1 \dots x_T}\left[-\sum_{t=1}^{T} x_t w_t\right] = -\sum_{t=1}^{T} \underbrace{\mathop{\mathbf{E}}_{x_t}[x_t]}_{0} w_t = 0$$

For any choices of $x_1 \dots x_T \in \{\pm 1\}$, the hypothesis $u \in \{\pm 1\}$ that achieves the smallest cumulative loss must minimize $-\sum_{t=1}^T x_t u$, which is either $-\sum_{t=1}^T x_t$ or $\sum_{t=1}^T x_t$. This implies that $u = \textbf{sign}\left(\sum_{t=1}^T x_t\right)$, with the optimal loss

$$l(u) = -\sum_{t=1}^T x_t \textbf{sign}\left(\sum_{t=1}^T x_t\right) = -\textbf{sign}\left(\sum_{t=1}^T x_t\right)\sum_{t=1}^T x_t = -\left|\sum_{t=1}^T x_t\right|$$

Thus for any $w_1 \dots w_T$, the expected regret is

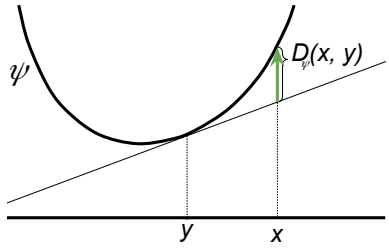
$$\mathop{\mathbf{E}}_{x_1 \dots x_T}\left[-\sum_{t=1}^T x_t w_t + \sum_{t=1}^T x_t \textbf{sign}\left(\sum_{t'=1}^T x_{t'}\right)\right] = 0 + \mathop{\mathbf{E}}_{x_1 \dots x_T}\left[\left|\sum_{t=1}^T x_t\right|\right] = \Theta(\sqrt{T})$$

The last term is $\Theta(\sqrt{T})$ just by the central limit theorem.[17] Thus we have constructed a randomized enemy that achieves an $\Omega(\sqrt{T})$ expected regret for any $w_1 \dots w_T$ asymptotically as $T \to \infty$. This implies the existence of *some* deterministic enemy that achieves an $\Omega(\sqrt{T})$ regret for any $w_1 \dots w_T$ asymptotically as $T \to \infty$ (aka. Yao's minimax principle). The intuition is that randomization is only a handicap for the enemy, not a feature.

C Bregman Divergence

Let $\psi : \Omega \to \mathbb{R}$ be a strictly convex and differentiable function over a convex set $\Omega \subseteq \mathbb{R}^d$. The associated **Bregman divergence** $D_\psi(x, y)$ (from y to x) measures the error of the first-order approximation of ψ around $y \in \Omega$ at $x \in \Omega$.

$$D_\psi(x, y) := \psi(x) - \psi(y) - \nabla\psi(y)^\top (x - y)$$



Since ψ is strictly convex, $D_\psi(x, y) \geq 0$ and zero iff $x = y$. We say ψ is **σ-strongly convex** with respect to the norm $||\cdot||$ if $D_\psi(x, y) \geq \frac{\sigma}{2}||x - y||^2$. $D_\psi(x, y)$ is clearly assymetric. It is trivially convex and differentiable in x with the gradient $\nabla_x D_\psi(x, y) = \nabla\psi(x) - \nabla\psi(y)$. It is not necessarily convex in y. The two most important examples of Bregman divergence are as follows:

1. For any $A \succ 0$, the A-weighted Euclidean norm $\psi : \mathbb{R}^d \to \mathbb{R}$ induces the **A-weighted Euclidean distance** (Appendix M.3).

$$\psi(x) = \frac{1}{2}||x||_A^2 \qquad \Rightarrow \qquad D_\psi(x, y) = \frac{1}{2}||x - y||_A^2 \tag{83}$$

   Clearly, ψ is 1-strongly convex wrt. $||\cdot||_A$. In particular, $\psi(x) = \frac{1}{2}||x||_2^2$ is 1-strongly convex wrt. the l_2 norm.

2. The negative entropy $\psi : \Delta^{d-1} \to \mathbb{R}$ induces the **KL divergence**.

$$\psi(x) = \sum_{i=1}^d x_i \log x_i \qquad \Rightarrow \qquad D_\psi(x, y) = \mathrm{KL}(x, y) \tag{84}$$

   Pinsker's inequality gives us $\mathrm{KL}(x, y) \geq \frac{1}{2}||x - y||_1^2$, thus ψ is 1-strongly convex wrt. the l_1 norm.

C.1 Generalized Pythagorean Theorem

Lemma C.1. For all $x, y, z \in \Omega$,

$$D_\psi(y, x) = D_\psi(z, x) + D_\psi(y, z) + (\nabla\psi(z) - \nabla\psi(x))^\top (y - z) \tag{85}$$

[17] Each x_t is an independent Rademacher variable with mean 0 and variance 1, so we have $|\sum_{t=1}^T x_t| \to \sqrt{T}\,|Z|$ where $Z \sim \mathcal{N}(0, 1)$. The boundedness of $|\cdot|$ implies $\mathbf{E}[|\sum_{t=1}^T x_t|] \to \sqrt{T}\mathbf{E}[|Z|]$ where $\mathbf{E}[|Z|] = \sqrt{2/\pi}$ is some constant.

Proof. Let $D_\psi(y,x) = D_\psi(z,x) + D_\psi(y,z) + C$ for some term C. Expanding by definition,

$$\begin{aligned} C &= D_\psi(y,x) - D_\psi(z,x) - D_\psi(y,z) \\ &= \{\cancel{\psi(y)} - \cancel{\psi(x)} - \nabla\psi(x)^\top(y - \cancel{x})\} - \{\cancel{\psi(z)} - \cancel{\psi(x)} - \nabla\psi(x)^\top(z - \cancel{x})\} - \{\cancel{\psi(y)} - \cancel{\psi(z)} - \nabla\psi(z)^\top(y - z)\} \\ &= (\nabla\psi(z) - \nabla\psi(x))^\top(y - z) \end{aligned}$$

□

Lemma C.2. Let $\mathcal{C} \subseteq \Omega$ be a convex and closed set. Pick any $x \in \Omega$. Let

$$p_x = \arg\min_{z \in \mathcal{C}} \; D_\psi(z,x) \tag{86}$$

denote the **Bregman projection** of $x \in \Omega$ onto $\mathcal{C}$. Then for all $y \in \mathcal{C}$,

$$D_\psi(y,x) \geq D_\psi(p_x,x) + D_\psi(y,p_x) \tag{87}$$

where the inequality is tight iff $\nabla\psi(p_x) - \nabla\psi(x)$ is orthogonal to $y - p_x$.

Proof. By Lemma C.1, we only need to show $(\nabla\psi(p_x) - \nabla\psi(x))^\top(y - p_x) \geq 0$ for all $y \in \mathcal{C}$. Since $p_x = \arg\min_{z\in\mathcal{C}} f(z)$ is the minimizer of the convex function $f(z) = D_\psi(z,x)$ over $\mathcal{C}$, it follows that $\nabla f(p_x)^\top(y - p_x) \geq 0$. But $\nabla f(p_x) = \nabla\psi(p_x) - \nabla\psi(x)$. □

Example C.1 (Pythagorean theorem)**.** Let $\Omega = \mathbb{R}^d$ and $\mathcal{C} \subseteq \mathbb{R}^d$ be a subspace. Let $\psi(x) = ||x||_2^2$ which induces the squared Euclidean distance $D_\psi(x,z) = ||x - z||_2^2$ in $\mathbb{R}^d$. Then $p_x \in \mathcal{C}$ is the orthogonal projection of x onto the subspace $\mathcal{C}$ where $x - p_x$ is orthogonal to $\mathcal{C}$. In particular, $\nabla\psi(p_x) - \nabla\psi(x) = p_x - x$ is orthogonal to $y - p_x$ for any $y \in \mathcal{C}$, hence (87) holds with equality, i.e., the usual Pythagorean theorem: $||x - y||_2^2 = ||x - p_x||_2^2 + ||p_x - y||_2^2$.

C.1.1 Regularized Bregman projection

We can further extend Lemma C.2 to regularize the Bregman projection with a convex function.

Lemma C.3. Let $\mathcal{C} \subseteq \Omega$ be a convex and closed set. Let $l : \mathcal{C} \to \mathbb{R}$ be convex and differentiable. Pick any $x \in \Omega$. Let

$$p_x = \arg\min_{z \in \mathcal{C}} \; D_\psi(z,x) + l(z) \tag{88}$$

denote the Bregman projection of $x \in \Omega$ onto $\mathcal{C}$ regularized by l. Then for all $y \in \mathcal{C}$,

$$D_\psi(y,x) + l(y) \geq D_\psi(p_x,x) + D_\psi(y,p_x) + l(p_x) \tag{89}$$

where the inequality is tight iff l is affine, i.e., $g = \nabla l(z)$ for any $z \in \mathcal{C}$, and $\nabla\psi(p_x) - \nabla\psi(x) + g$ is orthogonal to $y - p_x$.

Proof. Define $f(z) = D_\psi(z,x) + l(z)$ which is convex and differentiable. Since $p_x = \arg\min_{z\in\mathcal{C}} f(z)$, it follows that

$$\begin{aligned} 0 \leq \nabla f(p_x)^\top(y - p_x) &= (\nabla\psi(p_x) - \nabla\psi(x) + \nabla l(p_x))^\top(y - p_x) \\ &= D_\psi(y,x) - D_\psi(p_x,x) - D_\psi(y,p_x) + \nabla l(p_x)^\top(y - p_x) && \text{(Lemma C.1)} \\ &\leq D_\psi(y,x) - D_\psi(p_x,x) - D_\psi(y,p_x) + l(y) - l(p_x) \end{aligned}$$

where the inequality uses the convexity of l. Rearranging the terms gives (89). The second inequality is tight iff l is affine and the first inequality is tight iff $\nabla f(p_x)^\top(y - p_x) = 0$, thus (89) is tight iff both conditions hold. □

C.2 Other Properties

Lemma C.4 (Duality)**.** Assume Ω is closed. Then $D_\psi(y,x) = D_{\psi^*}(\nabla\psi(x), \nabla\psi(y))$

Proof. Since Ω is closed, $\psi^*(p) = \sup_{x\in\Omega} p^\top x - \psi(x)$ is well defined. Strict convexity implies $\nabla\psi : \Omega \to \mathbb{R}^d$ is invertible, so $\nabla(\psi^*) = (\nabla\psi)^{-1}$ and $\psi^*(p) = p^\top(\nabla\psi)^{-1}(p) - \psi((\nabla\psi)^{-1}(p))$ by Fact E.4. The latter implies that $\psi^*(\nabla\psi(z)) = \nabla\psi(z)^\top z - \psi(z)$. We can directly verify the equality:

$$\begin{aligned} D_{\psi^*}(\nabla\psi(x), \nabla\psi(y)) &= \psi^*(\nabla\psi(x)) - \psi^*(\nabla\psi(y)) - \nabla\psi^*(\nabla\psi(y))(\nabla\psi(x) - \nabla\psi(y)) \\ &= \left\{\nabla\psi(x)^\top x - \psi(x)\right\} - \left\{\nabla\psi(y)^\top y - \psi(y)\right\} - y^\top(\nabla\psi(x) - \nabla\psi(y)) \\ &= \psi(y) - \psi(x) - \psi(y)^\top(y - x) \\ &= D_\psi(y,x) \end{aligned}$$

□

Lemma C.5 (Mean as minimizer)**.** Let p be a distribution over a closed set $S \subseteq \Omega$. Define $x^\star = \arg\min_{x \in S} \mathbf{E}_{y\sim p}[D_\psi(y, x)]$. Then $x^\star = \mu_p$ (i.e., the mean of p).

Proof. By the linearity of expectation,

$$\mathop{\mathbf{E}}_{y\sim p}[D_\psi(y, x)] = \mathop{\mathbf{E}}_{y\sim p}[\psi(y)] - \psi(x) - \nabla\psi(x)^\top(\mu_p - x)$$

To see μ_p is optimal, consider any $x \in S$ and note

$$\mathop{\mathbf{E}}_{y\sim p}[D_\psi(y, x)] - \mathop{\mathbf{E}}_{y\sim p}[D_\psi(y, \mu_p)] = \psi(\mu_p) - \psi(x) - \nabla\psi(x)^\top(\mu_p - x) = D_\psi(\mu_p, x) \geq 0$$

which is minimized to zero at $x = \mu_p$. □

D Exponentiated Gradient Descent

Lemma D.1. In mirror descent (1), assume $d \geq 2$. Choose $V = \Delta^{d-1}$ and $\psi_t(w) = -H(w) = \sum_{i=1}^d w_i \log w_i$ so that the objective reduces to (see (84))

$$w_{t+1} = \mathop{\arg\min}_{w\in\Delta^{d-1}} g_t^\top w + \frac{1}{\eta_t}\mathrm{KL}(w, w_t) \tag{90}$$

where $w_t \in \Delta^{d-1}$ is assumed full-support. Then w_{t+1} satisfies

$$w_{t+1,i} = \frac{w_{t,i}\exp(-\eta_t g_{t,i})}{\sum_{j=1}^d w_{t,j}\exp(-\eta_t g_{t,j})} \tag{91}$$

Proof I. The objective is convex since KL is strictly convex in the first argument. The feasible set (probability simplex)

$$\Delta^{d-1} := \left\{ w \in \mathbb{R}^d : \; w \geq 0_d, \; \sum_{i=1}^d w_i = 1 \right\}$$

has only linear constraints with a strictly feasible point (since $d \geq 2$). Thus strong duality holds and we can solve the KKT system to find a global optimum. The Lagrangian is (omitting the nonnegativity constraint):

$$L(w, \lambda, \tau) = \eta_t g_t^\top w + \sum_{i=1}^d w_i \log\frac{w_i}{w_{t,i}} + \tau(1_d^\top w - 1)$$

The optimal solution satisfies the stationarity condition

$$\frac{\partial L(w, \lambda, \tau)}{\partial w_i} = \eta_t g_{t,i} + \log\frac{w_i}{w_{t,i}} + 1 + \tau = 0 \qquad \Leftrightarrow \qquad w_i = w_{t,i}\exp(-\eta_t g_{t,i} - 1 + \tau)$$

Enforcing the constraint $\sum_j w_j = 1$ yields $\tau = -\log(\sum_j w_{t,j}\exp(-\eta_t g_{t,j} - 1))$. Plugging it in the expression, we get a feasible solution:

$$w_i = \frac{w_{t,i}\exp(-\eta_t g_{t,i} - 1)}{\sum_j w_{t,j}\exp(-\eta_t g_{t,j} - 1)} = \frac{w_{t,i}\exp(-\eta_t g_{t,i})}{\sum_j w_{t,j}\exp(-\eta_t g_{t,j})} > 0$$

□

Proof II. The objective is equivalent to

$$\mathop{\mathbf{E}}_{i\sim w}\left[\eta_t g_{t,i} + \log\frac{w_i}{w_{t,i}}\right] = \mathop{\mathbf{E}}_{i\sim w}\left[\log\frac{w_i}{w_{t,i}\exp(-\eta_t g_{t,i})}\right] = \mathop{\mathbf{E}}_{i\sim w}\left[\log\frac{w_i}{u_{t,i}}\right] - \log z_t$$

where $z_t = \sum_j w_{t,j}\exp(-\eta_t g_{t,j})$ and $u_t = w_t/z_t \in \Delta^{d-1}$. Thus $w_{t+1} = \arg\min_{w\in\Delta^{d-1}} \mathrm{KL}(w, u) = u$. □

The argument in Proof II applies equally to the "KL-constrained RL problem" where the goal is to find the next policy by maximizing the expected reward $r(y) \in \mathbb{R}$ for action $y \sim \pi$ subject to the constraint $\mathrm{KL}(\pi, \pi_t) \leq C$.

$$\pi_{t+1} = \mathop{\arg\max}_{\pi\in\Delta^{d-1}} \mathop{\mathbf{E}}_{y\sim\pi}[r(y)] - \frac{1}{\eta_t}\mathrm{KL}(\pi, \pi_t) \qquad \Rightarrow \qquad \pi_{t+1}(y) \propto \pi_t(y)e^{r(y)}$$

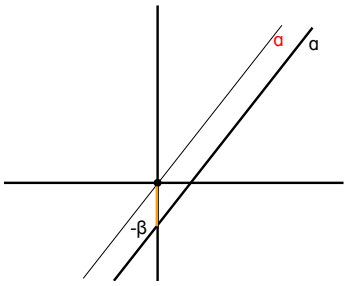
We can see that (90) is a special case where the action space is $i \in \{1 \dots d\}$ and the reward is the gradient $g_{t,i} \in \mathbb{R}$.

E Convex Conjugate

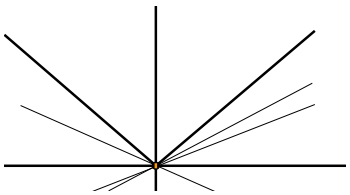
Let $\psi : \mathbb{R} \to \mathbb{R}$. Its **convex conjugate** $\psi^* : \mathbb{R} \to \mathbb{R}$ maps a slope p to how much px can overestimate $\psi(x)$:

$$\psi^*(p) = \max_{x \in \mathbb{R}} \left\{ px - \psi(x) \right\} \tag{92}$$

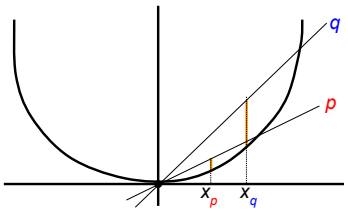
ψ^* is always convex no matter what ψ is (because $\psi^*(p)$ is the pointwise maximum of affine functions of p).



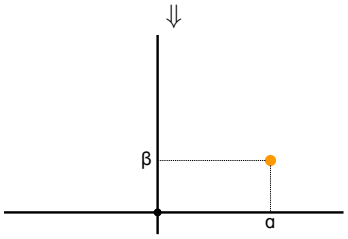
$\psi(x) = \alpha x - \beta$



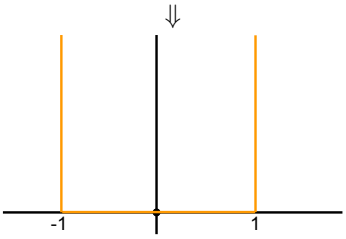
$\psi(x) = |x|$



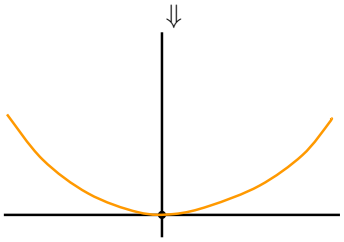
$\psi(x) = x^2$



$\psi^*(p) = \beta$ if $p = \alpha$, else ∞



$\psi^*(p) = 0$ if $|p| \leq 1$, else ∞



$\psi^*(p) = \frac{p^2}{4}$

Lemma E.1. If ψ is convex and differentiable with an invertible ψ', then $\psi^*(p) = p \times (\psi')^{-1}(p) - \psi((\psi')^{-1}(p))$.

Proof. Since ψ is convex, any $x_p \in \mathbb{R}$ satisfying $\psi'(x_p) = p$ is an optimal solution in (92). (This is visually clear in the rightmost example.) Since ψ' is invertible, $x_p = (\psi')^{-1}(p)$ is unique. □

Lemma E.2. If ψ is convex and differentiable with an invertible ψ', then ψ^* is differentiable with $(\psi^*)' = (\psi')^{-1}$.

Proof. By Lemma E.1, we have $\psi^*(p) = p \times (\psi')^{-1}(p) - \psi((\psi')^{-1}(p))$. An inverse function is differentiable, so we can use the product rule and the chain rule to obtain

$$(\psi^*)'(p) = (\psi')^{-1}(p) + p \times ((\psi')^{-1})'(p) - \underbrace{\psi'((\psi')^{-1}(p))}_{p} \times ((\psi')^{-1})'(p) = (\psi')^{-1}(p)$$

□

Lemma E.3. $\psi(x) \geq \psi^{**}(x)$ for all $x \in \mathbb{R}$. If ψ is convex and differentiable with an invertible ψ', then $\psi = \psi^{**}$.

Proof. For the first claim,

$$\begin{aligned} \psi^*(p) \geq px - \psi(x) \quad \forall x, p \in \mathbb{R} \quad &\Leftrightarrow \quad \psi(x) \geq px - \psi^*(p) \quad \forall x, p \in \mathbb{R} \\ &\Leftrightarrow \quad \psi(x) \geq \max_{p \in \mathbb{R}} \left\{ xp - \psi^*(p) \right\} = \psi^{**}(x) \quad \forall x \in \mathbb{R} \end{aligned}$$

For the second claim, since $\psi' : \mathbb{R} \to \mathbb{R}$ is a bijection,

$$\psi^{**}(x) = \max_{p \in \mathbb{R}} \left\{ xp - \psi^*(p) \right\} = \max_{y \in \mathbb{R}} \left\{ x\psi'(y) - \psi^*(\psi'(y)) \right\}$$

By Lemma E.1, the last term becomes

$$\psi^*(\psi'(y)) = \psi'(y) \times (\psi')^{-1}(\psi'(y)) - \psi((\psi')^{-1}(\psi'(y))) = \psi'(y)y - \psi(y)$$

Plugging this back in, we have

$$\psi^{**}(x) = \max_{y \in \mathbb{R}} \left\{ \psi(y) - \psi'(y)(y - x) \right\}$$

Using the fact that ψ is strongly convex (implied by the premise), we can easily verify that the RHS is maximized at $y = x$, thus $\psi^{**}(x) = \psi(x)$. Intuitively, the expression considers all lines tangent to ψ and picks the one that gives minimum underestimation at x. □

Exercise 1. Verify that Lemma E.1, E.2, and E.3 hold for $\psi(x) = x^2$.

E.1 Vector-Valued Input

The results for scalar-valued input easily generalize to vector-valued input $\psi : \mathbb{R}^d \to \mathbb{R}$. We summarize them below.

Fact E.4. Let $\psi : \mathbb{R}^d \to \mathbb{R}$. Its **convex conjugate** $\psi^* : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\psi^*(p) = \max_{x \in \mathbb{R}} \left\{ p^\top x - \psi(x) \right\} \tag{93}$$

ψ^* is convex and $\psi(x) \geq \psi^{**}(x)$ for all $x \in \mathbb{R}^d$. If ψ is convex and differentiable with an invertible gradient $\nabla\psi : \mathbb{R}^d \to \mathbb{R}^d$,

$$\psi^*(p) = p^\top (\nabla\psi)^{-1}(p) - \psi((\nabla\psi)^{-1}(p)) \tag{94}$$

$$\nabla(\psi^*) = (\nabla\psi)^{-1} \tag{95}$$

$$\psi = \psi^{**} \tag{96}$$

F Regularized Update Descent

To relate Polyak and Nesterov momentum, Botev *et al.* (2017) consider minimizing the objective $J(\theta) \in \mathbb{R}$ (assume θ is a scalar for simplicity) through

$$\widetilde{J}(\theta, v) := J(\theta + v) + \frac{\gamma}{2} v^2$$

where $v \in \mathbb{R}$ is a regularized velocity ($\gamma \in (0, 1)$). We can easily see that $\theta^\star, v^\star = \arg\min_{\theta, v} \widetilde{J}(\theta, v)$ satisfy $v^\star = 0$ and $\theta^\star = \arg\min_\theta J(\theta)$. They propose to do SGD wrt. the velocity v and then update the parameter $\theta \leftarrow \theta + v$ (instead of doing SGD wrt. θ). Choose initial θ_1, v_1 and for $t = 1, 2, \ldots$ compute

$$v_{t+1} = v_t - \eta_t \frac{\partial \widetilde{J}(\theta_t, v_t)}{\partial v} = \underbrace{(1 - \eta_t \gamma)}_{\mu_t} v_t - \eta_t J'(\theta_t + v_t) \tag{97}$$

followed by $\theta_{t+1} = \theta_t + v_{t+1}$. (97) clearly resembles SGD with momentum. Let $J^{(1)}_\theta(v) = J(\theta) + J'(\theta)v$ and $J^{(2)}_\theta(v) = J(\theta) + J'(\theta)v + \frac{\mu_t}{2} J''(\theta)v^2$ denote the first- and second-order approximations of $J(\theta + v)$ around θ, where the latter chooses to further regularize the second-order term with $\mu_t < 1$. Then

$$J'(\theta_t + v_t) \approx \frac{\partial J^{(1)}_{\theta_t}(v_t)}{\partial v} = J'(\theta_t)$$

$$J'(\theta_t + v_t) \approx \frac{\partial J^{(2)}_{\theta_t}(v_t)}{\partial v} = J'(\theta_t) + \mu_t J''(\theta_t) v_t \approx J'(\theta_t + \mu_t v_t)$$

where the latter again uses a first-order approximation. These yield

$$v_{t+1} = \mu_t v_t - \eta_t J'(\theta_t) \qquad \text{(Polyak)}$$

$$v_{t+1} = \mu_t v_t - \eta_t J'(\theta_t + \mu_t v_t) \qquad \text{(Nesterov)}$$

While the approach relates the two SGD momentum methods, it is a bit awkward to motivate (e.g., updateing the velocity instead of parameter does not guarantee that $J(\theta + v)$ is reduced). and not an exact match (e.g., $\mu_t \in (0, 1)$ must be scheduled and coupled with the learning rate).

G Kronecker Product

The **Kronecker product** $C = A \otimes B$ of $A \in \mathbb{R}^{m \times d}$ and $B \in \mathbb{R}^{n \times l}$ is defined as

$$C = \begin{bmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{bmatrix} \in \mathbb{R}^{mn \times dl}$$

(i.e., md copies of $B \in \mathbb{R}^{n \times l}$, each scaled by $A_{i,j} \in \mathbb{R}$). Specifically,

$$C_{(i_1-1)n+j_1,(i_2-1)l+j_2} = A_{i_1,i_2} \times B_{j_1,j_2} \tag{98}$$

for $i_1 \in [m]$, $j_1 \in [n]$, $i_2 \in [d]$, and $j_2 \in [l]$ (we shorthand $[N] = \{1 \dots N\}$). One way to see this is: for each row i_1 of A, we go through all the rows j_1 of B. Let $\overline{\text{vec}} : \mathbb{R}^{m \times n} \to \mathbb{R}^{mn}$ denote row-major vectorization (e.g., $\overline{\text{vec}}([[a,b];[c,d]]) = (a,b,c,d)$). Then (reference)

$$\overline{\text{vec}}(ABC) = (A \otimes C^\top)\overline{\text{vec}}(B) \tag{99}$$

(e.g., $\overline{\text{vec}}(uv^\top) = u \otimes v$). By the **mixed-product property**,

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{100}$$

Matrix transpose distributes without reordering as: $(A \otimes B)^\top = A^\top \otimes B^\top$.

G.1 Optimal Kronecker Decomposition

For any $A \in \mathbb{R}^{m \times d}$ and $B \in \mathbb{R}^{n \times l}$, we would like to define a permutation of the $mdnl$ values in $A \otimes B \in \mathbb{R}^{mn \times dl}$ into the shape $md \times nl$ such that

$$\textbf{rearrange}(A \otimes B) = \overline{\text{vec}}(A)\overline{\text{vec}}(B)^\top \tag{101}$$

We can reverse-engineer the correspondence. Since

$$(A \otimes B)_{(i_1-1)n+j_1,(i_2-1)l+j_2} = A_{i_1,i_2} \times B_{j_1,j_2} = (\overline{\text{vec}}(A)\overline{\text{vec}}(B)^\top)_{(i_1-1)d+i_2,(j_1-1)l+j_2}$$

for $i_1 \in [m]$, $j_1 \in [n]$, $i_2 \in [d]$, and $j_2 \in [l]$, we can define $\textbf{rearrange} : \mathbb{R}^{mn \times dl} \to \mathbb{R}^{md \times nl}$ by

$$\textbf{rearrange}(H)_{(i_1-1)d+i_2,(j_1-1)l+j_2} := H_{(i_1-1)n+j_1,(i_2-1)l+j_2} \tag{102}$$

We will assume that m, d, n, l are known when we call this function (for the given input of shape $mn \times dl$). By definition, (102) satisfies (101). A useful property of the function is that it is an involution (i.e., its own inverse):

$$\textbf{rearrange}(\textbf{rearrange}(H)) = H \tag{103}$$

One way to see this is to view the function simply as changing the way how we read the $mdnl$ input values by $(i_1, j_1, i_2, j_2) \mapsto (i_1, i_2, j_1, j_2)$ where we *swap* two axes, so applying it again recovers the original way. Van Loan and Pitsianis (1993) proposed the rearrangement for finding an optimal Kronecker decomposition of a matrix due to the following property:

Lemma G.1. Let $C \in \mathbb{R}^{mn \times dl}$. For any $A \in \mathbb{R}^{m \times d}$ and $B \in \mathbb{R}^{n \times l}$,

$$||C - A \otimes B||_F = \left|\left|\textbf{rearrange}(C) - \text{vec}(A)\text{vec}(B)^\top\right|\right|_F$$

Proof. Using (101), the obvious linearity of **rearrange**, and the fact that $||\cdot||_F$ is unaffected by rearranging values,

$$\begin{aligned} \left|\left|\textbf{rearrange}(C) - \text{vec}(A)\text{vec}(B)^\top\right|\right|_F &= ||\textbf{rearrange}(C) - \textbf{rearrange}(A \otimes B)||_F \\ &= ||\textbf{rearrange}(C - A \otimes B)||_F \\ &= ||C - A \otimes B||_F \end{aligned}$$

□

Corollary G.2. Let $C \in \mathbb{R}^{mn \times dl}$ and

$$A_\star, B_\star = \operatorname*{arg\,min}_{A \in \mathbb{R}^{m \times d}, B \in \mathbb{R}^{n \times l}} ||C - A \otimes B||_F$$

A solution is given by

$$\textbf{rearrange}(C) = \sum_i \sigma_i u_i v_i^\top \qquad A_\star = a \times \text{view}(u_1, m, d) \qquad \forall a, b > 0 : \; a \times b = \sigma_1$$
$$B_\star = b \times \text{view}(v_1, n, l)$$

where $M = \text{view}(u, d_1, d_2)$ arranges $u \in \mathbb{R}^{d_1 d_2}$ into a matrix of shape $M \in \mathbb{R}^{d_1 \times d_2}$ in row-major order (i.e., `view` in PyTorch). In other words, optimal Kronecker decomposition reduces to optimal rank-1 approximation of $\textbf{rearrange}(C) \in \mathbb{R}^{md \times nl}$, which is solvable by SVD.

G.2 Kronecker Product Between Square Matrices

If $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$ have eigenvalues $\lambda_1 \dots \lambda_m$ and $\mu_1 \dots \mu_n$, the mn eigenvalues of $A \otimes B \in \mathbb{R}^{mn \times mn}$ are $\lambda_1 \mu_1 \dots \lambda_m \mu_n$. It follows that

$$\text{tr}\,(A \otimes B) = \text{tr}\,(A)\,\text{tr}\,(B) \tag{104}$$
$$A, B \succeq 0 \quad \Rightarrow \quad (A \otimes B)^p = A^p \otimes B^p \quad \forall p \in \mathbb{R} \tag{105}$$
$$A, B \succeq 0 \quad \Rightarrow \quad A \otimes B \succeq 0 \tag{106}$$

From (106), we can also infer that[18]

$$A \succeq A' \succeq 0, \quad B \succeq B' \succeq 0 \qquad \Rightarrow \qquad A \otimes B \succeq A' \otimes B' \tag{107}$$

G.3 Outer Product Bound

Lemma G.3. Let $A \in \mathbb{R}^{m \times n}$ be any matrix where $\text{rank}\,(A) \le r$. If $a = \overline{\text{vec}}(A) \in \mathbb{R}^{mn}$,

$$aa^\top \preceq r(AA^\top) \otimes I_n$$
$$aa^\top \preceq r I_m \otimes (A^\top A)$$

Proof. Let $A = \sum_{k=1}^r \sigma_k u_k v_k^\top \in \mathbb{R}^{m \times n}$ be a thin SVD. Since $\overline{\text{vec}}$ is linear, $a = \sum_{k=1}^r \sigma_k \overline{\text{vec}}(u_k v_k^\top) = \sum_{k=1}^r \sigma_k\,(u_k \otimes v_k)$. Thus

$$aa^\top = \left(\sum_{k=1}^r \sigma_k\,(u_k \otimes v_k)\right)\left(\sum_{k=1}^r \sigma_k\,(u_k \otimes v_k)\right)^\top \preceq r \sum_{k=1}^r \sigma_k^2\,(u_k \otimes v_k)\,(u_k \otimes v_k)^\top \tag{108}$$
$$= r \sum_{k=1}^r \sigma_k^2 \left(u_k u_k^\top\right) \otimes \left(v_k v_k^\top\right)$$
$$\preceq r \sum_{k=1}^r \sigma_k^2 \left(u_k u_k^\top\right) \otimes I_n \tag{109}$$
$$= r(AA^\top) \otimes I_n$$

(108) uses the fact that $(\sum_{i=1}^r w_i)(\sum_{i=1}^r w_i)^\top \preceq r \sum_{i=1}^r w_i w_i^\top$ for any $w_1 \dots w_r \in \mathbb{R}^d$.[19] (109) follows from (107) since $I_n \succeq v_k v_k^\top$. □

We invoke the fact that the geometric mean of PSD matrices respects Loewner order (aka. "operator monotone"):

Fact G.4. Let $Y_1 \succeq X_1 \succeq 0$ and $Y_2 \succeq X_2 \succeq 0$ be PSD (square) matrices. Then $Y_1^\alpha Y_2^{1-\alpha} \succeq X_1^\alpha X_2^{1-\alpha}$ for all $\alpha \in [0, 1]$.

[18] $A \otimes B = (A' + C) \otimes (B' + D) = A' \otimes B' + A' \otimes D + C \otimes B' + C \otimes D \succeq A' \otimes B'$ since $C = A - A' \succeq 0$ and $D = B - B' \succeq 0$.

[19] This follows from Jensen's inequality and the convexity of $f(z) = z^2$ (i.e., $f(\sum_{i=1}^r z_i) \le (1/r) \sum_{i=1}^r f(z_i)$). Pick any $x \in \mathbb{R}^d$ and denote $z_i = x^\top w_i$. Then $x^\top \left(\sum_{i=1}^r w_i\right)\left(\sum_{i=1}^r w_i\right)^\top x = \left(\sum_{i=1}^r z_i\right)^2 \le r\left(\sum_{i=1}^r z_i^2\right) = r\left(\sum_{i=1}^r x^\top w_i w_i^\top x\right) = x^\top \left(r \sum_{i=1}^r w_i w_i^\top\right) x$.

Corollary G.5. Let $A \in \mathbb{R}^{m \times n}$ be any matrix where $\operatorname{rank}(A) \le r$. If $a = \overline{\operatorname{vec}}(A) \in \mathbb{R}^{mn}$,

$$aa^\top \preceq r(AA^\top \otimes A^\top A)^{1/2} \tag{110}$$

Proof. By Lemma G.3, we have $r(AA^\top) \otimes I_n \succeq aa^\top$ and $rI_m \otimes (A^\top A) \succeq aa^\top$. Applying Fact G.4, we have $aa^\top \preceq r((AA^\top) \otimes I_n)(I_m \otimes (A^\top A)) = r(AA^\top \otimes A^\top A)$. □

H Hessian

Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ denote an input-label pair. Let $f_w : \mathcal{X} \to \mathbb{R}^K$ denote a neural network parameterized by $w \in \mathbb{R}^d$. Let $L : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ denote a loss function differentiable in the first argument. The most important loss is the cross-entropy loss with $\mathcal{Y} = \{1 \dots K\}$ given by $L(z, y) = -\log p_z(y) = \log(\sum_k e^{z_k}) - z_y$, whose gradient is famously $\nabla_z L(z, y) = p_z - e_y \in \mathbb{R}^K$ where $e_y \in \{0, 1\}^K$ is the y-th standard basis. We typically rely on the gradient

$$\nabla_w L(f_w(x), y) = \frac{\partial L(f_w(x), y)}{\partial w} \in \mathbb{R}^d \tag{111}$$

for optimizing f_w. In contrast, the Hessian

$$H_{x,y}(w) = \nabla_w^2 L(f_w(x), y) = \frac{\partial^2 L(f_w(x), y)}{\partial w^2} \in \mathbb{R}^{d \times d} \tag{112}$$

is avoided because of the d^2 size, even though it yields a much faster convergence rate (e.g., Newton's method). It is informative to analyze (112) nonetheless. As in backpropagation, we first decompose it by disentangling f_w and L via the chain rule (Appendix I). This yields

$$H_{x,y}(w) = \underbrace{\nabla_w f_w(x)}_{d \times K} \underbrace{\left(\nabla_z^2 L(z, y)\big|_{z=f_w(x)}\right)}_{K \times K} \underbrace{\nabla_w f_w(x)^\top}_{K \times d} + \underbrace{\nabla_w^2 f_w(x)}_{d \times d \times K} \underbrace{\left(\nabla_z L(z, y)\big|_{z=f_w(x)}\right)}_{K \times 1} \tag{113}$$

The first "outer" term, involving only the $d \times K$ Jacobian of f_w and the $K \times K$ Hessian of L, is called the **Gauss-Newton (GN) component** of the Hessian. GN is empirically found to be a good approximation of the Hessian (Sankar *et al.*, 2021); it is exact if f_w is linear (since the second term vanishes). Given a population distribution **pop** over (x, y) and assuming the cross-entropy loss, we can further relate GN with the gradient (111).

Lemma H.1. Let $L(z, y) = -\log p_z(y)$. Then

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}}\left[\nabla_w f_w(x)\left(\nabla_z^2 L(z, y)\big|_{z=f_w(x)}\right)\nabla_w f_w(x)^\top\right] = \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}\\ \hat{y}\sim f_w(x)}}\left[\nabla_w L(f_w(x), \hat{y})\nabla_w L(f_w(x), \hat{y})^\top\right] = I(w) \tag{114}$$

The RHS of (114) coincides with the Fisher information matrix $I(w)$ (i.e., the covariance of $\nabla_w L(f_w(x), y)$ where $y \sim f_w(x)$).[20]

Proof of Lemma H.1. Note that regardless of the label $y \in \{1 \dots K\}$, the Jacobian of the cross-entropy loss is the same as the Jacobian of the softmax function:

$$\nabla_z^2 L(z, y) = \nabla_z(p_z - e_y) = \nabla_z p_z = \operatorname{diag}(p_z) - p_z p_z^\top$$

Using the fact that $\mathbf{E}[e_{\hat{y}}] = p_z$ and $\mathbf{E}[e_{\hat{y}} e_{\hat{y}}^\top] = \operatorname{diag}(p_z)$ where $\hat{y} \sim p_z$, we can express this as a vector outer product:

$$\mathop{\mathbf{E}}_{\hat{y}\sim p_z}\left[(p_z - e_{\hat{y}})(p_z - e_{\hat{y}})^\top\right] = \operatorname{diag}(p_z) - p_z p_z^\top$$

Putting together, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{(x,y)\sim\mathbf{pop}}\left[\nabla_w f_w(x)\left(\nabla_z^2 L(z, y)\big|_{z=f_w(x)}\right)\nabla_w f_w(x)^\top\right] &= \mathop{\mathbf{E}}_{x\sim\mathbf{pop}}\left[\nabla_w f_w(x)\left(\operatorname{diag}(p_{f_w(x)}) - p_{f_w(x)} p_{f_w(x)}^\top\right)\nabla_w f_w(x)^\top\right] \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}\\ \hat{y}\sim f_w(x)}}\left[\nabla_w f_w(x)(p_{f_w(x)} - e_{\hat{y}})(p_{f_w(x)} - e_{\hat{y}})^\top \nabla_w f_w(x)^\top\right] \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}\\ \hat{y}\sim f_w(x)}}\left[\nabla_w L(f_w(x), \hat{y})\nabla_w L(f_w(x), \hat{y})^\top\right]
\end{aligned}$$

where the last equality is the chain rule: $\nabla_w L(f_w(x), \hat{y}) = \nabla_w f_w(x)(\nabla_z L(z, \hat{y})\big|_{z=f_w(x)})$. □

[20] The expected Hessian $H(w) = E_{(x,y)\sim\mathbf{pop}}[H_{x,y}(w)] = I(w) + E_{(x,y)\sim\mathbf{pop}}[\nabla_w^2 f_w(x)(\nabla_z L(z, y)\big|_{z=f_w(x)})]$ is still not exactly Fisher since the second term does not vanish in general. Nonetheless, many works on second-order optimization assume $H(w) \approx I(w)$.

I Vector Calculus Scratch Pad

Let $w \in \mathbb{R}^d$. We can verify

$$\nabla_w g(f(w)) = (\nabla_w f(w))(\nabla_{f(w)} g(f(w))) \qquad \forall f : \mathbb{R}^d \to \mathbb{R}^M,\ g : \mathbb{R}^M \to \mathbb{R}^K \qquad \textbf{(chain rule)}$$
$$\nabla_w (F(w)g(w)) = (\nabla_w F(w))g(w) + F(w)(\nabla_w g(w))^\top \qquad \forall g : \mathbb{R}^D \to \mathbb{R}^K,\ F : \mathbb{R}^d \to \mathbb{R}^{D\times K} \qquad \textbf{(product rule)}$$

where $\nabla_w F(w) \in \mathbb{R}^{d\times D\times K}$ is the Jacobian of $F(w) \in \mathbb{R}^{D\times K}$. Let $z = z(w) \in \mathbb{R}^K$ (activations) and $L = L(z) \in \mathbb{R}$ (loss). We write

$$\nabla_w L \in \mathbb{R}^d : (\nabla_w L)_i = \frac{\partial L}{\partial w_i} \qquad \text{(gradient of the loss wrt. the weights)}$$
$$\nabla_z L \in \mathbb{R}^K : (\nabla_z L)_k = \frac{\partial L}{\partial z_k} \qquad \text{(gradient of the loss wrt. the activations)}$$
$$\nabla_w z \in \mathbb{R}^{d\times K} : (\nabla_w z)_{i,k} = \frac{\partial z_k}{\partial w_i} \qquad \text{(Jacobian of the activations wrt. the weights)}$$
$$\nabla_w^2 L \in \mathbb{R}^{d\times d} : (\nabla_w^2 L)_{i,j} = \frac{\partial^2 L}{\partial w_i \partial w_j} \qquad \text{(Hessian of the loss wrt. the weights)}$$
$$\nabla_z^2 L \in \mathbb{R}^{K\times K} : (\nabla_z^2 L)_{k,l} = \frac{\partial^2 L}{\partial z_k \partial z_l} \qquad \text{(Hessian of the loss wrt. the activations)}$$
$$\nabla_w^2 z \in \mathbb{R}^{d\times d\times K} : (\nabla_w^2 z)_{i,j,k} = \frac{\partial^2 z_k}{\partial w_i \partial w_j} \qquad \text{(Hessians of the activations wrt. the weights)}$$

By the chain rule, we have

$$\nabla_w L = (\nabla_w z)(\nabla_z L)$$
$$\nabla_w (\nabla_z L) = (\nabla_w z)(\nabla_z^2 L)$$

By the product rule, we have

$$\nabla_w^2 L = (\nabla_w^2 z)(\nabla_z L) + (\nabla_w z)(\nabla_z^2 L)(\nabla_w z)^\top$$

J Integral Form of the Taylor Expansion

We can write any $J : \mathbb{R}^d \to \mathbb{R}$ as

$$J(w) = J(u) + \nabla J(u)^\top (w-u) + \int_0^1 (1-t)(w-u)^\top \nabla^2 J(u + t(w-u))(w-u)dt \tag{115}$$

This "lossless" Taylor expansion only uses up to the second-order derivatives. It is a direct consequence of the FTC which allows us to write $f : \mathbb{R} \to \mathbb{R}$ as (Lemma S.4)

$$f(1) = f(0) + f'(0) + \int_0^1 (1-t) f''(t) dt$$

Using $f(t) = J(u + t(w-u))$ (i.e., path of J from u to w) yields (115). It has applications where we want to express functional chracteristics in terms of its Hessian.

Lemma J.1. $J : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex wrt. the l_2 norm (i.e., $J(w) \geq J(u) + \nabla J(u)^\top (w-u) + \frac{m}{2} ||w-u||_2^2$) iff $\nabla^2 J(w) \succeq m I_d$.

Proof. If $\nabla^2 J(x) \succeq m I_d$, then $v^\top \nabla^2 J(x) v \geq m ||v||_2^2$ for all $v \in \mathbb{R}^d$, so from (115)

$$\begin{aligned} J(w) &= J(u) + \nabla J(u)^\top (w-u) + \int_0^1 (1-t)(w-u)^\top \nabla^2 J(u + t(w-u))(w-u)dt \\ &\geq J(u) + \nabla J(u)^\top (w-u) + \left(m ||w-u||_2^2\right) \int_0^1 (1-t)dt \\ &= J(u) + \nabla J(u)^\top (w-u) + \frac{m}{2} ||w-u||_2^2 \end{aligned} \tag{116}$$

Conversely starting from (116) with $w \neq u$, writing $w = u + tv$ for some $t > 0$ and $v \in \mathbb{R}^d$ we have

$$\frac{J(u+tv) - J(u) - t\nabla J(u)^\top v}{t^2/2} \geq m\,||v||_2^2 \tag{117}$$

Define $\phi(t) = J(u+tv)$ where $\phi(t) = \phi(0) + \phi'(0)t + \frac{1}{2}\phi''(0)t^2 + o(t^2)$. Observe $\phi(0) = J(u)$, $\phi'(0) = \nabla J(u)^\top v$, and $\phi''(0) = v^\top \nabla^2 J(u) v$. Since (117) holds for any $t > 0$, this implies $v^\top \nabla^2 J(u) v \geq m\,||v||_2^2$. □

Lemma J.2. $J(w) \leq J(u) + \nabla J(u)^\top (w-u) + \frac{M}{2}\,||w-u||_2^2$ iff $\nabla^2 J(w) \preceq M I_d$.

Proof. The proof is symmetric to that of Lemma J.1. If J is convex, the former is equivalent to M-smoothness and the latter is equivalent to $\left|\left|\nabla^2 J(w)\right|\right|_2 \leq M$, consistently with Lemma S.18. □

In a similar vein, the integral form is useful when we want to express the gradient in terms of the "mean" Hessian. Let $w^\star$ denote any stationary point of $J : \mathbb{R}^d \to \mathbb{R}$. The multivariate version of the FTC allows us to write

$$\begin{aligned}
\nabla J(w) &= \nabla J(w^\star) + \int_0^1 \nabla^2 J(w^\star + t(w-w^\star))(w-w^\star)dt \\
&= \left(\int_0^1 \nabla^2 J(w^\star + t(w-w^\star))dt\right)(w-w^\star) \\
&= \bar{H}(w^\star, w)(w-w^\star)
\end{aligned} \tag{118}$$

where $\bar{H}(w^\star, w) \in \mathbb{R}^{d\times d}$ integrates the Hessian from $w^\star$ to w.

K Root Mean Square (RMS)

In the context of measuring "typical" per-element size of some $x \in \mathbb{R}^d$, we often use the **RMS** (root mean square)

$$\mathrm{RMS}(x) = \sqrt{\frac{1}{d}\sum_{i=1}^d x_i^2} = \frac{||x||_2}{\sqrt{d}}$$

When $d = 1$, this becomes $\sqrt{x^2} = |x|$, so it should be thought of as a multi-dimensional generalization of absolute value. The average inside the square root is a second-moment estimator if $x_1 \dots x_d$ are iid as X, i.e.,

$$\mathbf{E}[X^2] = \lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^d x_i^2$$

Thus a natural theoretical counterpart to RMS is $\sqrt{\mathbf{E}[X^2]}$ where X represents any element in $x \in \mathbb{R}^d$. When $\mathbf{E}[X] = 0$, it coincides with standard deviation which is a standard way to measure the "size" of a random variable; otherwise, it is simply the square root of the second moment.

Why not directly work with absolute value? A more direct approach is to take the average of absolute values and consider its theoretical counterpart

$$\mathrm{MeanAbs}(x) = \frac{1}{d}\sum_{i=1}^d |x_i| \qquad\qquad \mathbf{E}\,|X| = \lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^d |x_i|$$

The reason we prefer RMS is mainly mathematical convenience. Analyzing the second moment of X is relatively easy (e.g., known properties of variance, linearity of expectation) while analyzing the absolute value is nonstandard and requires special tools. Note that RMS is always an upper bound on the absolute value by Cauchy-Schwarz $\mathbf{E}\,|X| \leq \sqrt{\mathbf{E}[X^2]}$. In some cases, the absolute value is given analytically and can be related to RMS. For instance if $X \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbf{E}\,|X| = \sigma\sqrt{2/\pi}$ and thus

$$\sqrt{\mathbf{E}[X^2]} = \sqrt{\frac{\pi}{2}}\mathbf{E}\,|X| \approx 1.25\mathbf{E}\,|X|$$

which shows RMS is equivalent to absolute value up to a constant factor.

What about the mean? Keep in mind that the mean says absolutely nothing about typical *size*. Consider $x \in \mathbb{R}^{100}$ where $x_i = -42$ for $i = 1 \dots 50$ and $x_i = 42$ for $i = 51 \dots 100$. Then

$$\frac{1}{100}\sum_{i=1}^{100} x_i = 0 \qquad\qquad \mathrm{RMS}(x) = 42$$

We see that RMS correctly captures the "absolute value" of x. In contrast, the mean captures the average "drift". The theoretical counterpart is X such that $\mathbf{E}[X] = 0$ but $\mathrm{Var}\,(X) \gg 0$.

L Nonnegative Matrix Factorization (NMF)

Let $C \in \mathbb{R}_{\geq 0}^{m \times n}$ and $r \in \mathbb{N}$. We wish to find $A \in \mathbb{R}_{\geq 0}^{m \times r}$ and $B \in \mathbb{R}_{\geq 0}^{r \times n}$ such that $C \approx AB$. A natural divergence between nonnegative values is the I-divergence (Finesso and Spreij, 2006; Lee and Seung, 1999). For any $a, b \geq 0$, the I-divergence is defined as

$$\mathrm{IDiv}(a, b) = a \log \frac{a}{b} - a + b \tag{119}$$

where $\frac{0}{0} = 0$ and $0 \log 0 = 0$. (119) is nonnegative due to the convexity of $x \log x$. For multi-dimensional inputs p, q (of the same shape), we define $\mathrm{IDiv}(p, q) = \sum_{i=1}^{n} \mathrm{IDiv}(p_i, q_i)$. In particular, $\mathrm{IDiv}(p, q) = \mathrm{KL}(p, q)$ if p, q are distributions. Minimizing $\mathrm{IDiv}(C, AB)$ over nonnegative A, B is equivalent to

$$A^\star, B^\star = \operatorname*{arg\,min}_{A \in \mathbb{R}_{\geq 0}^{m \times r}, B \in \mathbb{R}_{\geq 0}^{r \times n}} \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n} -C_{i,j} \log\left(\sum_{k=1}^{r} A_{i,k}B_{k,j}\right) + \sum_{k=1}^{r} A_{i,k}B_{k,j}}_{J_C(A,B)} \tag{120}$$

We have a manifold of optimal solutions $J_C(A^\star, B^\star) = J_C(\alpha A^\star, \frac{1}{\alpha} B^\star)$. The objective is biconvex. Putting aside the nonnegative constraints for now, the gradient is given by (we focus on A since B is analogous):

$$\frac{\partial J_C(A, B)}{\partial A_{i,k}} = -\frac{\sum_{j=1}^{n} C_{i,j}B_{k,j}}{\sum_{l=1}^{r} A_{i,l}B_{l,j}} + \sum_{j=1}^{n} B_{k,j}$$

In general, there is no closed-form solution for a stationary point. We can still do projected gradient descent on A, but a more popular approach is the multiplicative update (123) which preserves nonnegativity.

Rank one. If $r = 1$, the stationary point has a closed-form solution.[21]

$$A_{i,1} = -\frac{\sum_{j=1}^{n} C_{i,j}}{A_{i,1}} + \sum_{j=1}^{n} B_{1,j} = 0 \qquad \Leftrightarrow \qquad A_{i,1} = \frac{\sum_{j=1}^{n} C_{i,j}}{\sum_{j=1}^{n} B_{1,j}} \qquad \Leftrightarrow \qquad A = \frac{C1_n}{B1_n}$$

Similarly, we have the stationary $B = \frac{1_m^\top C}{1_m^\top A}$. We may constrain $A \in \mathbb{R}^{m \times 1}$ to satisfy $1_m^\top A = 1_m^\top C 1_n$ (using scale invariance) so that $A = C1_n$ and $B = \frac{1_m^\top C}{1_m^\top C 1_n}$. Since they are nonnegative, they are a solution to (120). Thus rank-one NMF is easy (even though it is still technically nonconvex).

L.1 A Generative Story

We assume a model parameterized by $A \in \mathbb{R}_{\geq 0}^{m \times r}$ and $B \in \mathbb{R}_{\geq 0}^{r \times n}$. It generates the latent variable $Z \in \mathbb{N}_0^{m \times n \times r}$ by

$$Z_{i,j,k} \sim \mathrm{Poi}(A_{i,k}B_{kj})$$

Then it generates the observation $C \in \mathbb{N}_0^{m \times n}$ by $C_{i,j} = \sum_{k=1}^{r} Z_{i,j,k}$. Since $C_{i,j} \sim \mathrm{Poi}(\sum_{k=1}^{r} A_{i,k}B_{kj})$ by the usual property of Poisson, the marginal distribution over C is

$$p_{A,B}(C) = \prod_{i,j} \frac{(\sum_{k=1}^{r} A_{i,k}B_{kj})^{C_{i,j}} e^{-\sum_{k=1}^{r} A_{i,k}B_{kj}}}{C_{i,j}!}$$

[21] From the generative perspective of the next section, this happens largely because we remove the "latent variable" and the "sum inside log".

The joint distribution over Z and C satisfying $C_{i,j} = \sum_{k=1}^{r} Z_{i,j,k}$ is

$$p_{A,B}(Z, C) = \prod_{i,j,k} \frac{(A_{i,k}B_{kj})^{Z_{i,j,k}} e^{-A_{i,k}B_{kj}}}{Z_{i,j,k}!}$$

The posterior over $Z_{i,j} \in \mathbb{N}_0^r$ conditioned on $C_{i,j}$ follows the multinomial distribution (Lemma S.2):

$$p_{A,B}(Z_{i,j}|C_{i,j}) = \text{Mult}\left(C_{i,j}, \left(\frac{A_{i,k}B_{kj}}{\sum_l A_{i,l}B_{lj}}\right)_{k=1}^{r}\right)(Z_{i,j}) \tag{121}$$

We seek the MLE, i.e., the maximizer of the marginal log-likelihood:

$$\begin{aligned} A^\star, B^\star &= \operatorname*{arg\,max}_{A \in \mathbb{R}_{\geq 0}^{m \times r}, B \in \mathbb{R}_{\geq 0}^{r \times n}} \log p_{A,B}(C) \\ &= \operatorname*{arg\,min}_{A \in \mathbb{R}_{\geq 0}^{m \times r}, B \in \mathbb{R}_{\geq 0}^{r \times n}} \sum_{i=1}^{m}\sum_{j=1}^{n} -C_{i,j} \log\left(\sum_{k=1}^{r} A_{i,k}B_{kj}\right) + \sum_{k=1}^{r} A_{i,k}B_{kj} \end{aligned} \tag{122}$$

We see that (122) and (120) are the same. But now that we have a generative story, we can do EM. At any A, B, we can maximize the ELBO using the exact posterior (121) to find

$$\begin{aligned} A', B' &= \operatorname*{arg\,max}_{\widehat{A} \in \mathbb{R}_{\geq 0}^{m \times r}, \widehat{B} \in \mathbb{R}_{\geq 0}^{r \times n}} \mathop{\mathbf{E}}_{Z \sim p_{A,B}(\cdot|C)} \left[\log p_{\widehat{A},\widehat{B}}(Z, C)\right] \\ &= \operatorname*{arg\,max}_{\widehat{A} \in \mathbb{R}_{\geq 0}^{m \times r}, \widehat{B} \in \mathbb{R}_{\geq 0}^{r \times n}} \sum_{i,j,k} C_{i,j}\left(\frac{A_{i,k}B_{kj}}{\sum_l A_{i,l}B_{lj}}\right) \log(\widehat{A}_{i,k}\widehat{B}_{k,j}) - \widehat{A}_{i,k}\widehat{B}_{k,j} \end{aligned}$$

As usual with EM, the sum inside log is moved outside. Solving the stationary condition, we have the blockwise update[22]

$$A'_{i,k} = A_{i,k} \times \left(\frac{\sum_j C_{i,j}B_{k,j}}{\sum_l A_{i,l}B_{lj}}\right) / \left(\sum_j B_{k,j}\right) \qquad B'_{k,j} = B_{k,j} \times \left(\frac{\sum_i C_{i,j}A_{i,k}}{\sum_l A_{i,l}B_{lj}}\right) / \left(\sum_i A_{i,k}\right) \tag{123}$$

Note that the multiplicative update preserves nonnegativity (assuming A, B are nonnegative). In matrix form, the update is

$$R = C \oslash AB \qquad \begin{aligned} A' &= A \odot (RB^\top \oslash 1_n^\top B^\top) \\ B' &= B \odot (AR^\top \oslash A^\top 1_m) \end{aligned} \tag{124}$$

where $\oslash$ is elementwise division (broadcasted) and $\odot$ is elementwise multiplication.

L.2 AdamNMF

Using (124), we can easily motivate a rank-r generalization of Adafactor (Section 4.7). A pseudocode is given below. The memory overhead in estimating the second gradient moment is $O((m+n)r)$ as opposed to $O(mn)$ in Adam.

AdamNMF

Input: initial layer weight $W_1 \in \mathbb{R}^{m \times n}$, rank $r \geq 1$, learning rate $\eta > 0$, initialization range $\epsilon > 0$

1. $A_0 \sim \text{Unif}(0, \epsilon)^{m \times r}$, $B_0 \sim \text{Unif}(0, \epsilon)^{r \times n}$
2. For $t = 1 \ldots T$:
   (a) Receive the gradient $G_t \in \mathbb{R}^{m \times r}$, compute the elementwise square G_t^2.
   (b) Do one round of EM to decompose $G_t^2 \approx A_t B_t$ using A_{t-1} and B_{t-1} as initialization:

$$R_{t-1} \leftarrow G_t^2 \oslash A_{t-1}B_{t-1} \qquad \begin{aligned} A_t &\leftarrow A_{t-1} \odot \left(R_{t-1}B_{t-1}^\top \oslash 1_n^\top B_{t-1}^\top\right) \\ B_t &\leftarrow B_{t-1} \odot \left(A_{t-1}R_{t-1}^\top \oslash A_{t-1}^\top 1_m\right) \end{aligned}$$

   (c) $W_{t+1} \leftarrow W_t - \eta \frac{G_t}{\sqrt{A_t B_t}}$
3. Return $W_{T+1} \in \mathbb{R}^{m \times n}$

[22] There's a bit more going on here, since we update one variable while holding the other fixed. This version of EM is so-called "Generalized EM". It simply means breaking the M-step into sub-updates for blocks of parameters; as long as the sub-updates do not decrease the MLL, the convergence property of EM remains.

M Vector Spaces

M.1 Normed Spaces

The function $||\cdot|| : \mathcal{V} \to \mathbb{R}_{\geq 0}$ is a **norm** on a vector space $\mathcal{V}$ if it satisfies (i) the triangle inequality $||u+v|| \leq ||u||+||v||$, (ii) the *absolute* homogeneity $||\alpha u|| = |\alpha| \cdot ||u||$, and (iii) the point-separating property $||u|| = 0 \Rightarrow u = 0_d$. For $\mathcal{V} = \mathbb{R}^d$, a broad family of norms is given by the l_p-norm:

$$||w||_p := \left(\sum_{i=1}^{d} |w_i|^p \right)^{1/p} \qquad \forall p \geq 1$$

This includes the popular l_2, l_1, l_∞ norms:

$$||w||_2 = \sqrt{\sum_i w_i^2} \qquad \text{(Euclidean)}$$

$$||w||_1 = \sum_i |w_i| \qquad \text{(taxicab)}$$

$$||w||_\infty := \lim_{p \to \infty} ||w||_p = \max_i |w_i| \qquad \text{(maximum)}$$

On the other hand, the "l_0 norm" defined as $||w||_0 := |\{i : w_i \neq 0\}|$ is often mentioned in the context of promoting sparsity, but it is not a norm (e.g., violates the traingle inequality).

M.2 Inner Product Spaces

The function $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is an **inner product** on the (real) vector space $\mathcal{V}$ if it is symmetric, linear in the first argument, and positive-definite (i.e., $\langle u, u \rangle \geq 0$ with equality iff u is zero). An inner product induces a **canonical norm** by $||u|| = \sqrt{\langle u, u \rangle}$, thus an inner product space always a normed space. The most important inner product on $\mathcal{V} = \mathbb{R}^d$ is the **dot product** $\langle u, v \rangle = u^\top v = \sum_i u_i v_i$ which induces the l_2 norm. In contrast, there is no inner product that induces the l_1 or l_∞ norm.

M.2.1 Dual norm

For any norm $||w||$, the **dual norm** $||\cdot||_* : \mathcal{V} \to \mathbb{R}_{\geq 0}$ is defined as[23]

$$||v||_* := \sup_{w \in \mathcal{V}:\ ||w|| \leq 1} w^\top v \tag{125}$$

The definition arises naturally in an effort to bound the dot product since

$$w^\top v \leq ||w||\, ||v||_* \tag{126}$$

for all $v, w \in \mathcal{V}$. (126) is referred to as "generalized Cauchy-Schwarz" or more accurately (the finite-dimensional version of) **Hölder's inequality**. It can be verified that the dual norm is a norm itself and an involution (i.e., $||w||_{**} = ||w||$).

M.3 Weighted Euclidean Norm

For any $d \times d$ positive-definite matrix $A \succ 0$, we define a "weighted Euclidean norm" by

$$||u||_A := \left|\left| A^{1/2} u \right|\right|_2 = \sqrt{u^\top A u} \tag{127}$$

[23] This is a *different* definition of the dual norm from Hilbert spaces (i.e., inner product spaces in infinite dimensions). There, the dual norm is defined as $||v||_* := \sup_{w \in \mathcal{V}:\ ||w|| \leq 1} \langle w, v \rangle$. One can verify that $||v||_* = ||v||$ ("self-dual") using the standard Cauchy-Schwarz inequality $|\langle u, v \rangle| \leq ||u||\, ||v||$. (The Cauchy-Schwarz inequality can be proved directly without dual norms, so there is no circular argument here.)

We can directly check that $||\cdot||_A$ is a norm on $\mathcal{V} = \mathbb{R}^d$.[24] To derive the dual norm, we observe

$$\begin{aligned}
||v||_* &= \max_{w \in \mathbb{R}^d:\, w^\top A w = 1} w^\top v \\
&= \max_{u \in \mathbb{R}^d:\, u^\top u = 1} u^\top A^{-1/2} v && (u = A^{1/2} w) \\
&\le \max_{u \in \mathbb{R}^d:\, u^\top u = 1} ||u||_2 \left|\left| A^{-1/2} v \right|\right|_2 && \text{(Cauchy-Schwarz)} \\
&= \sqrt{v^\top A^{-1} v} \\
&= ||v||_{A^{-1}}
\end{aligned}$$

Choosing $u \propto A^{-1/2} v$ yields a solution that makes the bound tight, thus $||v||_* = ||v||_{A^{-1}}$. See this note for a proof using the method of Lagrangian multipliers.

M.3.1 General $A \succeq 0$

Let $A \succeq 0$ with $r = \operatorname{rank}(A) \le d$. Let $A = V \Lambda V^\top$ denote a thin eigendecomposition where $V \in \mathbb{R}^{d \times r}$ is an orthonormal basis of $\operatorname{range}(A)$ and $\Lambda = \operatorname{diag}(\lambda_1 \dots \lambda_r)$ for $\lambda_i > 0$. Pick any $w \in \operatorname{range}(A)$. Then $w = Vx$ for some nonzero $x \in \mathbb{R}^r$, so that

$$w^\top A w = x^\top V^\top V \Lambda V^\top V x = x^\top \Lambda x > 0$$

Thus $||u||_A = \sqrt{u^\top A u}$ is a norm on $\mathcal{V} = \operatorname{range}(A)$. To derive the dual norm, we can take similar steps:

$$\begin{aligned}
||v||_* &= \max_{w \in \mathbb{R}^d:\, w^\top A w = 1} w^\top v \\
&= \max_{x \in \mathbb{R}^r:\, x^\top \Lambda x = 1} x^\top V^\top v \\
&= \max_{u \in \mathbb{R}^r:\, u^\top u = 1} u^\top \Lambda^{-1/2} V^\top v && (u = \Lambda^{1/2} x) \\
&\le \max_{u \in \mathbb{R}^d:\, u^\top u = 1} ||u||_2 \left|\left| \Lambda^{-1/2} V^\top v \right|\right|_2 && \text{(Cauchy-Schwarz)} \\
&= \sqrt{v^\top V \Lambda^{-1} V v} \\
&= \sqrt{v^\top A^+ v} \\
&= ||v||_{A^+}
\end{aligned}$$

We can again verify that choosing $u \propto \Lambda^{-1/2} V^\top v$ achieves this bound, thus $||v||_* = ||v||_{A^+}$. This subsumes the above analysis when $r = d$. When $r < d$ (i.e., A is rank-deficient), we have $\mathbb{R}^d = \operatorname{range}(A) \perp \operatorname{null}(A)$ (since A is symmetric) with a nontrivial null space. In particular, there exist nonzero $w \in \operatorname{null}(A)$ such that $||w||_A = \sqrt{w^\top A w} = 0$, thus $||\cdot||_A$ fails to satisfy the point-separating property on $\operatorname{null}(A)$ (i.e., it becomes a "seminorm" on $\mathcal{V} = \mathbb{R}^d$). Pick any nonzero $v \in \operatorname{null}(A)$. Assuming $r > 0$, we can select some $w_0 \in \operatorname{range}(A)$ such that $w_0^\top A w_0 = 1$. Define $w(\alpha) = w_0 + \alpha v$ and note that $w(\alpha)^\top A w(\alpha) = 1$ for all $\alpha \in \mathbb{R}$. Thus

$$\begin{aligned}
||v||_* &= \max_{w \in \mathbb{R}^d:\, w^\top A w = 1} w^\top v \\
&\ge \max_{\alpha \in \mathbb{R}} w(\alpha)^\top v \\
&= \max_{\alpha \in \mathbb{R}} w_0^\top v + \alpha ||v||_2^2 \\
&= \infty
\end{aligned}$$

(i.e., $||v||_*$ is not finite for $v \notin \operatorname{range}(A)$.)

N Intuitions for the Fundamental Theorem of Calculus

Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable with an integrable f'. Imagine finely chopping $[a, b]$ into $t_1 < \dots < t_{n+1}$. Clearly, the difference between $f(a)$ and $f(b)$ can be exactly "gathered" as

$$f(b) - f(a) = \sum_{i=1}^{n} \Big(f(t_{i+1}) - f(t_i) \Big)$$

[24] We can also view $||u||_A$ as the canonical norm of the inner product $\langle u, v \rangle_A := u^\top A v$ on $\mathcal{V} = \mathbb{R}^d$.

By the mean value theorem, there is some $\xi_i \in (t_i, t_{i+1})$ such that $f(t_{i+1}) - f(t_i) = f'(\xi_i)\Delta t_i$ ("slope $\times$ width"), so

$$f(b) - f(a) = \sum_{i=1}^{n} f'(\xi_i)\Delta t_i$$

Taking the limit $n \to \infty$ gives the "Part II" of the FTC:

$$f(b) - f(a) = \int_a^b f'(t)dt \tag{128}$$

N.1 Relative Knob

We typically use a relative knob $r \in [0, 1]$. We finely chop $[0, 1]$ into $r_1 < \cdots < r_{n+1}$ and gather the differences

$$f(b) - f(a) = \sum_{i=1}^{n} \Big(f(a + r_{i+1}(b-a)) - f(a + r_i(b-a)) \Big)$$

Note the width of a slice is now $(b-a)\Delta r_i$. Again by the mean value theorem, for some $\eta_i \in (r_i, r_{i+1})$

$$f(b) - f(a) = \sum_{i=1}^{n} f'(a + \eta_i(b-a))(b-a)\Delta r_i$$

Taking the limit $n \to \infty$

$$f(b) - f(a) = \int_0^1 f'(a + r(b-a))(b-a)dr \tag{129}$$

To check this algebraically, define the 1D path $\phi(r) = f(a+r(b-a))$ and apply (128) using $f(b)-f(a) = \phi(1)-\phi(0)$.

N.2 Higher Dimensions

The locally linear functional change remains beautifully invariant to dimensionality:

$f : \mathbb{R}^d \to \mathbb{R}$	$df \approx \langle \nabla f(x), dx \rangle$	("directional slope $\times$ displacement")
$F : \mathbb{R}^d \to \mathbb{R}^m$	$dF \approx J_F(x)dx$	("Jacobian $\times$ displacement")

One difference is that the displacement requires a choice of path since there are multiple ways go from a to b. A natural choice is the "straight" path $\gamma(r) = a + r(b-a)$ in $\mathbb{R}^d$ with displacement $\gamma'(r) = b - a \in \mathbb{R}^d$. Thus

$$f(b) - f(a) = \int_0^1 \langle \nabla f(\gamma(r)), \gamma'(r) \rangle \, dr = \int_0^1 \langle \nabla f(a + r(b-a)), b-a \rangle \, dr \tag{130}$$

$$F(b) - F(a) = \int_0^1 J_F(\gamma(r))\gamma'(r)dr = \int_0^1 J_F(a + r(b-a))(b-a)dr \tag{131}$$

N.3 Part I

The "Part I" of the FTC is an identity statement $A'_{x_0}(x) = f(x)$ where $A_{x_0}(x) = \int_{x_0}^{x} f(t)dt$ an antiderivative of (continuous) f, interpreted as signed area from a "base point" x_0.[25] This follows from the rectangular approximation

$$A_{x_0}(x + h) - A_{x_0}(x) \approx f(x)h$$

As the slice gets thinner, the approximation becomes increasingly accurate. This implies

$$f(x) = \lim_{h \to 0} \frac{A_{x_0}(x + h) - A_{x_0}(x)}{h} = A'_{x_0}(x)$$

[25] Thus f has many antiderivatives each differing by a constant $A(x) = A_{x_0}(x) + C$. Let x_1 be the base point of A. Then $A_{x_0}(x) - A_{x_1}(x) = \int_{x_0}^{x_1} f(t)dt$ is constant in x.

O Smoothness

There are a few ways to assert that $f : \mathbb{R}^d \to \mathbb{R}$ does not "change too much". Clearly, we first need to choose some norm $||\cdot|| : \mathbb{R}^d \to \mathbb{R}$. Common examples in the literature are

$$\left|\left|\nabla^2 f(x)\right|\right|_{||\cdot||\to||\cdot||_*} \leq L \tag{132}$$

$$||\nabla f(x) - \nabla f(z)||_* \leq L\,||x - z|| \tag{133}$$

$$|f(x) - f(z) - \langle \nabla f(z), x - z\rangle| \leq \frac{L}{2}\,||x - z||^2 \tag{134}$$

where $L > 0$ is a smoothness constant. (132) upper bounds the operator norm of the Hessian, defined as

$$\left|\left|\nabla^2 f(x)\right|\right|_{||\cdot||\to||\cdot||_*} := \sup_{h\in\mathbb{R}^d:||h||=1} \left|\left|\nabla^2 f(x)h\right|\right|_* \tag{135}$$

which collapses to the familiar spectral norm $\left|\left|\nabla^2 f(x)\right|\right|_2 = \max_i |\lambda_i|$ when $||\cdot|| = ||\cdot||_2$ since $\nabla^2 f(x)$ is symmetric (though not necessarily PSD) and $||\cdot||_* = ||\cdot||_2$. (133) is so-called L-Lipschitz continuous gradient. (134) upper bounds the linearization error.[26] Note that we measure all "gradient-like" quantities in the dual norm $||\cdot||_*$ because it is defined to make the pairing bound $|\langle \cdot, x\rangle| \leq ||\cdot||_*\,||x||$ hold. In general (Lemma S.17):

$$(132) \Leftrightarrow (133) \Rightarrow (134)$$

For the choice of Euclidean norm $||\cdot|| = ||\cdot||_2$, they are all equivalent (Lemma S.18).

O.1 Contrast with Convexity

Strong convexity assumes that

$$f(x) - f(z) - \langle \nabla f(z), x - z\rangle \geq \frac{m}{2}\,||x - z||^2 \tag{136}$$

for some $m > 0$. When $||\cdot|| = ||\cdot||_2$, this is equivalent to a lower bound on the minimum eigenvalue of the Hessian $\lambda_{\min}(\nabla^2 f(x)) \geq m$ (Lemma J.1). While (136) is seemingly similar to (134), it is a drastically stronger assumption compared to smoothness and makes the minimizer unique (when it exists).

P Majorization-Minimization Principle

Majorization-minimization (MM) refers to a broad design principle for minimizing a function $f : \mathcal{X} \to \mathbb{R}$. At the current iterate $z \in \mathcal{X}$, we build a surrogate $U(\cdot; z)$ such that

$$U(z; z) = f(z) \qquad \text{("touch")} \tag{137}$$

$$U(x; z) \geq f(x) \;\; \forall x \in \mathcal{X} \qquad \text{("majorize")} \tag{138}$$

We iterates for $k = 1, 2, \ldots$

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} U(x; x_k) \tag{139}$$

By design, it monotonically decreases the objective value since

$$f(x_k) = U(x_k; x_k) \geq U(x_{k+1}; x_k) \geq f(x_{k+1}) \tag{140}$$

Convergence to a stationary point can be shown under additional conditions. A natrual surrogate is the linearization of f at z with squared norm regularization:

$$U(x; z) = f(z) + \langle \nabla f(z), x - z\rangle + \frac{1}{2\eta}\,||x - z||^2$$

It is visually clear in Euclidean space that (137) is satisfied and (138) is also satisfied if the chosen regularization dominates linearization error. It is trivial to verify that if f is L-smooth (any of 132-134 in Appendix O), then

[26] $D_f(x, z) = f(x) - f(z) - \langle \nabla f(z), x - z\rangle$ is called the Bregman divergence when f is convex, becoming squared Euclidean distance if f is the squared l_2 norm and KL divergence if f is the negative entropy (within the probability simplex) (Appendix C). But clearly it can be used to measure the degree of nonlinearity for nonconvex f.

$U(x,z) \geq f(x)$ for $\eta \leq \frac{1}{L}$. That is, the smoother f is (smaller L), the less we need to regularize (smaller $1/(2\eta)$) to ensure U is an upper bound. In unconstrained Euclidean space (i.e., $\mathcal{X} = \mathbb{R}^d$, $||\cdot|| = ||\cdot||_2$), the iterates (139) are analytically given by

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

which is gradient descent. Indeed, the dynamic in (140) is mechanically how gradient descent "works".

Q Matrix Iteration

Q.1 Power Iteration

For simplicity let $X = V\Lambda V^\top = \sum_{i=1}^n \lambda_i v_i v_i^\top \in \mathbb{R}^{n\times n}$ be a symmetric PSD matrix with nonnegative eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ and corresponding orthonormal eigenvectors $v_1 \dots v_n \in \mathbb{R}^n$. Pick a random $\hat{v}_0 \in \mathbb{R}^n$ from the unit sphere and consider for some power $P \in \mathbb{N}$

$$\hat{v}_P = X^P \hat{v}_0 = \sum_{i=1}^n \lambda_i^P a_i v_i = \lambda_1^P a_1 \left(v_1 + r_P\right) \tag{141}$$

where we abbreviate $a_i = v_i^\top \hat{v}_0 \in \mathbb{R}$ (we use the fact that $a_1 \neq 0$ almost surely). The error vector $r_P \in \mathbb{R}^n$ takes the form $r_P = \sum_{i>1} (\lambda_i/\lambda_1)^P (a_i/a_1) v_i$. In particular, $||r_P||_2^2 = O((\lambda_2/\lambda_1)^{2P})$ and (141) converges as $\hat{v}_P / ||\hat{v}_P||_2 \to \mathbf{sign}\,(a_1)\, v_1$ if $\lambda_1 > \lambda_2$.[27] But the scale $\lambda_1^P a_1$ may blow up if $\lambda_1 \gg 1$ or vanish if $\lambda_1 \ll 1$, so in practice we iterate

$$q_{k+1} = \frac{X q_k}{||X q_k||_2}$$

with $q_0 = \hat{v}_0$. We can easily check that $q_k = \hat{v}_k / ||\hat{v}_k||_2$ at every step (induction), so the theory remains the same.

Q.2 Orthogonal Iteration

We may pursue the top-m eigenspace $\text{span}(V_m)$ where $V_m = [v_1 \dots v_m]$ by picking a random orthonormal $\widehat{V}_0 = [\hat{v}_{0,1} \dots \hat{v}_{0,m}] \in \mathbb{R}^{n\times m}$ from the Stiefel manifold and iterating

$$\widetilde{V}_{k+1} = X \widehat{V}_k \tag{142}$$

$$\widehat{V}_{k+1} = \text{QR}(\widetilde{V}_{k+1}) \tag{143}$$

where $\widehat{V}_{k+1}$ is an orthonormal basis of the column space of $\widetilde{V}_{k+1}$ computed exactly by QR factorization. Clearly, without the QR step this is simply doing m power iterations in parallel which would cause all m vectors to converge directionally to v_1. QR re-orthonormalizes the same column space and acts like implicit deflation. To draw a parallel with the power iteration analysis, we may express $\widehat{V}_k = V_m B_k + V_\perp C_k$ and note

$$\widetilde{V}_{k+1} = X \widehat{V}_k = V_m \Lambda_m B_k + V_\perp \Lambda_\perp C_k$$

Since the top part gets multiplied by Λ_m and the leakage part by $\Lambda_\perp$, the latter relatively shrinks like λ_{m+1}/λ_m. One can show that if $\lambda_m > \lambda_{m+1}$ and $V_m^\top \widehat{V}_0 \in \mathbb{R}^{m\times m}$ is invertible (holds almost surely), the orthogonal iteration satisfies $\text{span}(\widehat{V}_k) \to \text{span}(V_m)$, with the subspace error decaying as $\sin\Theta(\widehat{V}_k, V_m) = O((\lambda_{m+1}/\lambda_m)^k)$.[28]

Compute. A Householder-based QR on an $n \times m$ matrix ($n \geq m$) explicitly forming Q performs roughly $4nm^2 - (4/3)m^3$ FLOPs ($(8/3)n^3$ for a square matrix). Thus K orthogonal iterations (142–143) cost $K(2n^2m + 4nm^2 - 4/3m^3)$ FLOPs ($(14/3)n^3$ for a square matrix). In comparison, an exact SVD typically performs bidiagonalization (again by Householder) and a specialized bidiagonal SVD routine, costing around $4nm^2 - (4/3)m^3$ plus additional $O(nm^2 + m^3)$ work. In practice, it is a few times more expensive than QR.

[27] If no gap, it will directionally converge to the normalized projection of $\hat{v}_0$ onto the top eigensubspace.

[28] For full basis tracking (i.e., $m = n$), the convergence rate is dominated by $\max_i(\lambda_{i+1}/\lambda_i)^k$ since each column i converges at a different rate $(\lambda_{i+1}/\lambda_i)^k$. In contrast, tracking only the top-m subspace is governed by the single boundary ratio λ_{m+1}/λ_m.

Q.2.1 Online version

We often have a situation where $X^{(1)}, X^{(2)}, \ldots \in \mathbb{R}^{n\times n}$ drifts and want to track its basis, which is identifiable up to sign flips, permutations, and rotations (within a degenerate eigenspace). Naively we would perform exact eigendecomposition $X^{(t)} = V^{(t)}\Lambda^{(t)}(V^{(t)})^\top$ every step. Instead, power iteration allows for a natural online algorithm where we initialize $\widehat{V}^{(1)} \leftarrow V^{(1)}$ exactly and for $t = 2, 3\ldots$ compute

$$\widetilde{V}^{(t)} = X^{(t)}\widehat{V}^{(t-1)} \tag{144}$$

$$\widehat{V}^{(t)} = \mathrm{QR}(\widetilde{V}^{(t)}) \tag{145}$$

(i.e., single iteration on the previous estimate). This works if $X^{(t)} \approx X^{(t-1)}$ (e.g., EMA of the gradient second moment) and the eigengaps are not close to zero (otherwise the associated eigenvectors are not uniquely determined). Even so, each iterate typically needs permutation and sign corrections to achieve consistent dimension labeling. The sign consistency can be enforced easily, for instance flip the sign of $\hat{v}_i^{(t)}$ if $\langle \hat{v}_i^{(t)}, \hat{v}_i^{(t-1)}\rangle < 0$. The ordering consistency can be enforced by sorting the approximate eigenvalues $\hat{\lambda}_i^{(t)} = \hat{v}_i^{(t)} X^{(t)} \hat{v}_i^{(t)}$ by which the columns of $\widehat{V}^{(t)}$ are reordered, or more simply (though less accelerator-friendly) match the previous column ordering by $\max_j |\langle \hat{v}_i^{(t)}, \hat{v}_j^{(t-1)}\rangle|$.

Q.3 Newton-Schulz Iteration

Newton-Schulz is an application of a general contraction-based principle: design a function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$\phi(\phi(\phi(\cdots\phi(x)\cdots))) \to 1 \qquad\qquad \forall 0 < x \le 1 \tag{146}$$

For instance, a sufficient set of conditions is

$$\begin{aligned} 0 < \phi(x) \le 1 \qquad & \forall 0 < x \le 1 \\ \phi(x) > x \qquad & \forall 0 < x < 1 \\ \phi(1) = 1 \qquad & \end{aligned}$$

(e.g., concave on $[0, 1]$ with maximum of 1 at $x = 1$). It is visually clear the rate of convergence depends on how "flat" ϕ is near $x = 1$, since it implies $x_{\text{in}} \lesssim 1$ gets sent immediately close to the optimum $\phi(x_{\text{in}}) \approx 1$.[29]

$$\phi(x) = \phi(1) + \phi'(1)(x-1) + \frac{\phi''(1)}{2}(x-1)^2 + \frac{\phi'''(1)}{6}(x-1)^3 + \cdots$$

Let $\epsilon = 1 - x \in [0, 1)$ denote the error of the current point and $\epsilon_{\text{next}} = 1 - \phi(x) < \epsilon$ the next. Since $\phi(1) = 1$, rearranging gives us

$$\epsilon_{\text{next}} = \phi'(1)\epsilon - \frac{\phi''(1)}{2}\epsilon^2 + \frac{\phi'''(1)}{6}\epsilon^3 - \cdots = O(\epsilon^{n_{\text{loc}}})$$

where $n_{\text{loc}} \ge 1$ is the location of the first nonzero term (the higher, the flatter).[30] Hence after K iterations of $x_{k+1} = \phi(x_k)$ starting from $x_0 \in (0, 1]$, the error shrinks quadratically (i.e., doubly exponentially) in n_{loc} as $\epsilon_K = O(\mathrm{pow}(\epsilon_0, n_{\text{loc}}^K))$ once sufficiently small. Note that convergence collapses to $\epsilon_K = O(\phi'(1)^K\epsilon_0)$ if $n_{\text{loc}} = 1$ (i.e., only linear convergence assuming $|\phi'(1)| < 1$), so we typically assert $\phi'(1) = 0$.

Q.3.1 Application to matrix orthogonalization

Let $X = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ be the SVD of a matrix. The "orthogonalized" quantity $UV^\top \in \mathbb{R}^{m\times n}$ is a projection of X onto orthonormal matrices:

$$O^\star = \min_{O:\, OO^\top = I_{m\times m}} ||X - O||_F = \max_{O:\, ||O||_2 = 1} \langle X, O\rangle = UV^\top$$

In lieu of (146), it is natural to numerically estimate $\widehat{O} \approx O^\star$ by first setting $X_0 = X/||X||_F = U\Sigma_0 V^\top$ to make the singular values in range $\Sigma_{0,i,i} \in (0, 1]$[31] and iterating, for some $P \ge 2$ and $c_1 \ldots c_P \in \mathbb{R}$,

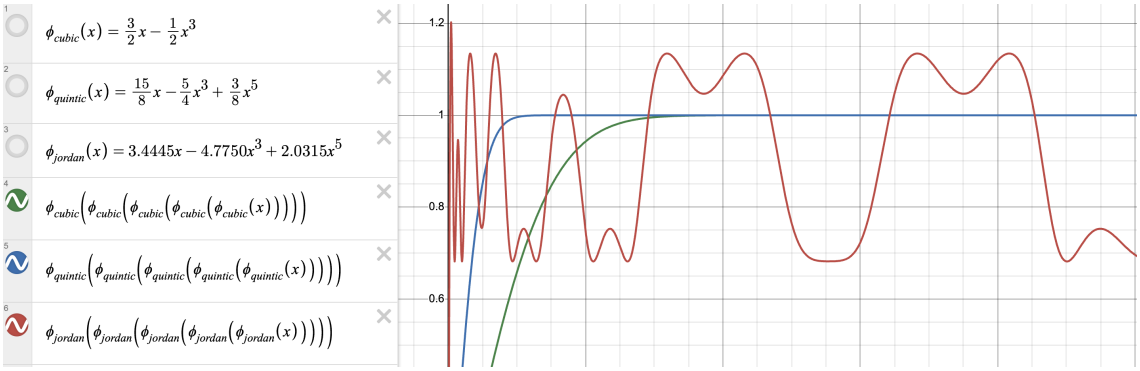
$$\begin{aligned} X_{k+1} &= c_1 X_k + c_2(X_k X_k^\top)X_k + c_3(X_k X_k^\top)^2 X_k + \cdots + c_P(X_k X_k^\top)^{P-1} X_k \\ &= U\left(c_1\Sigma_k + c_2\Sigma_k^3 + c_3\Sigma_k^5 + \cdots + c_P\Sigma_k^{2P-1}\right)V^\top \end{aligned} \tag{147}$$

[29] Note also that this is the opposite behavior of gradient-based optimization, which slows down on flat regions.
[30] It must be finite unless $\phi = 1$ (in which case there is no work to be done).
[31] It would be more proper to normalize by the spectral norm, but it is less computationally convenient.

This is equivalent to applying (146) to each normalized singular value $x = \Sigma_{0,i,i} \in (0, 1]$ using an *odd* polynomial

$$\phi(x) = c_1 x + c_2 x^3 + c_3 x^5 + \cdots + c_P x^{2P-1} = \sum_{i=1}^{P} c_i x^{2i-1}$$

We choose $P \geq 2$ and find $c_1 \dots c_P$ such that $\phi(1) = \sum_{i=1}^{P} c_i = 1$ (stationarity) and $\phi'(1) = 0$ (and possibly higher derivatives) to ensure at least quadratic local convergence. Choosing a higher P allows us to push n_{loc} but makes the iteration more costly. $P = 2$ or $P = 3$ is a good compromise. For $\phi(x) = c_1 x + c_2 x^3$, we can solve the 2×2 linear system $c_1 + c_2 = 1$ and $c_1 + 3c_2 = 0$ to obtain the unique solution $(c_1, c_2) = (3/2, -1/2)$. For $\phi(x) = c_1 x + c_2 x^3 + c_3 x^5$, we can solve the 3×3 linear system $c_1 + c_2 + c_3 = 1$, $c_1 + 3c_2 + 5c_3 = 0$, and $6c_2 + 20c_3 = 0$ to obtain the unique solution $(c_1, c_2, c_3) = (15/8, -5/4, 3/8)$. Jordan *et al.* (2024) propose $(c_1, c_2, c_3) = (3.4445, -4.7750, 2.0315)$ to rapidly flatten all singular values to $[1 \pm 0.3]$ in $K = 5$ iterations. It is best to see a visualization:



Compute. Following Jordan *et al.* (2024), an economical way to compute (147) for $X_k \in \mathbb{R}^{m \times n}$ with $P = 3$ is

$$\begin{aligned} A_k &= X_k X_k^\top && (2m^2 n \text{ FLOPs}) \\ B_k &= c_2 A_k + c_3 A_k A_k && (2m^3 \text{ FLOPs}) \\ X_{k+1} &= c_1 X_k + B_k X_k && (2m^2 n \text{ FLOPs}) \end{aligned}$$

for a total of $30m^3$ FLOPs assuming $K = 5$ and $m \approx n$. In comparison, doing 5 iterations of power iteration would cost $\approx 24m^3$ FLOPs (only $\approx 5m^3$ if online) (Appendix Q.2), While nominally of similar asymptotic order, power iteration is less practical compared to NS (which is pure matmul and thus especially accelerator-friendly). Furthermore, the convergence of NS does not depend on adjacent eigengaps but on the singular value spread.

R Subgradients

Let $f : \mathcal{V} \to \mathbb{R}$ denote a convex function over a finite-dimensional inner product space. A **subgradient** of f at $z \in \mathcal{V}$ is any $g \in \mathcal{V}$ such that the corresponding linearization lower bounds f:

$$f(x) \geq f(z) + \langle g, x - z \rangle \qquad \forall x \in \mathcal{V} \tag{148}$$

The set of all such subgradients is denoted by $\partial f(z)$. It is visually clear that if f is differentiable at z we must have the singleton $\partial f(z) = \{\nabla f(z)\}$.

R.1 Linear Chain Rule

Let $L : \mathcal{X} \to \mathcal{V}$ is any linear operator. Let $\phi : \mathcal{X} \to \mathbb{R}$ denote an associated affine-convex composite function $\phi(x) = f(L(x) + b)$ where $b \in \mathcal{V}$. The **linear chain rule for subgradients** is

$$\partial \phi(x) = L^*(\partial f(L(x) + b)) \tag{149}$$

where $L^* : \mathcal{V} \to \mathcal{X}$ is the adjoint of L characterized by $\langle y, L(x) \rangle = \langle L^*(y), x \rangle$. We only prove the easy direction $\partial \phi(x) \supseteq L^*(\partial f(L(x) + b))$ for intuition as the other direction is more complicated.

Proof. Pick any $g \in \partial f(L(z)+b)$ at some $z \in \mathcal{X}$. (148) implies for all $x \in \mathcal{X}$

$$\begin{aligned} f(L(x)+b) &\geq f(L(z)+b) + \langle g, (L(x)+b) - (L(z)+b)\rangle \\ &= f(L(z)+b) + \langle g, L(x-z)\rangle \\ &= f(L(z)+b) + \langle L^*(g), x-z\rangle \end{aligned}$$

This is equivalent to the statement: $\phi(x) \geq \phi(z) + \langle L^*(g), x-z\rangle$ for all $x \in \mathcal{X}$. Thus $L^*(g) \in \partial\phi(z)$. □

R.2 Subgradients of Norms

A norm $||\cdot|| : \mathcal{V} \to \mathbb{R}$ is always convex. We give a few practical characterizations of the subgradients of a vector/matrix norm composed with linear transformation, applying the linear chain rule (149):

- ($\mathcal{V} = \mathbb{R}^d$) Define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(\lambda) = ||\lambda x + b||$ at some $x, b \in \mathbb{R}^d$. Then

$$\partial\phi(\lambda) = \left\{x^\top z : \ z \in \partial\,||\lambda x + b||\right\} \tag{150}$$

  Proof: The underlying linear operator $L : \mathbb{R} \to \mathbb{R}^d$ is $L(\lambda) = \lambda x$. Its adjoint $L^* : \mathbb{R}^d \to \mathbb{R}$ is $L^*(z) = x^\top z$ since $\langle z, L(\lambda)\rangle = \lambda x^\top z = \langle L^*(z), \lambda\rangle$.

- ($\mathcal{V} = \mathbb{R}^{D\times d}$) Define $\phi : \mathbb{R}^{d\times d} \to \mathbb{R}$ by $\phi(\Lambda) = ||X\Lambda + B||$ at some $X, B \in \mathbb{R}^{D\times d}$. Then

$$\partial\phi(\Lambda) = \left\{X^\top Z : \ Z \in \partial\,||X\Lambda + B||\right\} \tag{151}$$

  Proof: The underlying linear operator $L : \mathbb{R}^{d\times d} \to \mathbb{R}^{D\times d}$ is $L(\Lambda) = X\Lambda$. Its adjoint $L^* : \mathbb{R}^{D\times d} \to \mathbb{R}^{d\times d}$ is $L^*(Z) = X^\top Z$ since $\langle Z, L(\Lambda)\rangle = \operatorname{tr}\left(Z^\top X\Lambda\right) = \left\langle X^\top Z, \Lambda\right\rangle$.

Another useful characterization is given below. The proof is left as an exercise.

$$\partial\,||x|| = \{z \in \mathcal{V} : \ ||z||_* \leq 1, \ \langle z, x\rangle = ||x||\} \tag{152}$$

S Lemmas

Lemma S.1. Let $O_T = \boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{d\times d}$ where $\boldsymbol{G} = (g_1 \ldots g_T) \in \mathbb{R}^{T\times d}$ is the matrix of gradients with rank d. Then

$$\operatorname{tr}\left(O_T^{1/2}\right) \leq \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2} \tag{153}$$

with equality iff O_T is diagonal; more specifically, $O_T = \operatorname{diag}\left(\sigma_1^2 \ldots \sigma_d^2\right)$ where $\sigma_1 > \ldots > \sigma_d > 0$ are the (distinct, for convenience) singular values of $\boldsymbol{G}$.

Proof. Let $\boldsymbol{G} = U\Sigma V^\top$ denote an SVD of $\boldsymbol{G}$. Then $O_T = V\Sigma^2 V^\top$ and thus $O_T^{1/2} = V\Sigma V^\top$, so the LHS can be expressed as $\operatorname{tr}\left(O_T^{1/2}\right) = \sum_{i=1}^d \sigma_i$ (i.e., the nuclear norm $||\boldsymbol{G}||_{\text{nuc}}$). Let $\gamma_1 \ldots \gamma_d \in \mathbb{R}^T$ denote the columns of $\boldsymbol{G}$. Note that $\gamma_i = \boldsymbol{G}e_i = U\Sigma\tilde{v}_i$ where $\tilde{v}_i \in \mathbb{R}^d$ is the i-th row of $V \in \mathbb{R}^{d\times d}$ (thus $\tilde{v}_1 \ldots \tilde{v}_d$ are orthonormal). Then the RHS can be expressed as $\sum_{i=1}^d ||\gamma_i||_2 = \sum_{i=1}^d ||U\Sigma\tilde{v}_i||_2 = \sum_{i=1}^d ||\Sigma\tilde{v}_i||_2$. Thus the claim (153) can be rephrased as: given any matrix with singular values Σ and right singular vectors V containing rows $\tilde{v}_1 \ldots \tilde{v}_d \in \mathbb{R}^d$, we must always ahve

$$\sum_{i=1}^{d} ||\Sigma e_i||_2 \leq \sum_{i=1}^{d} ||\Sigma\tilde{v}_i||_2 \tag{154}$$

Since Σ is on both sides, we vary the choice of V. WLOG we can assume that V is a 2×2 rotation matrix for the following reasons:

- V is orthonormal. So it can be expressed as a product of Givens rotations and at most one reflection (i.e., $\operatorname{diag}(-e_i)\, I_d$).
- Reflection does not affect the RHS of (154).

- Thus if no rotation in a 2D subspace reduces the RHS of (154), neither does any V.

Hence assuming $\tilde{v}_1 = (\cos\theta, -\sin\theta)$ and $\tilde{v}_2 = (\sin\theta, \cos\theta)$ for some radian θ, we can write down the objective:

$$\min_{0\le\theta<2\pi} \sqrt{\sigma_1^2\cos^2\theta + \sigma_2^2\sin^2\theta} + \sqrt{\sigma_1^2\sin^2\theta + \sigma_2^2\cos^2\theta}$$

We can easily check that the minimum is $\sigma_1+\sigma_2$ and the minimizers are $\theta^\star \in \left\{0, \frac{\pi}{2}\right\}$ corresponding to $V = I_2$ and $V = [[0,1],[1,0]]$. The latter violates the structure of V imposed by the SVD (i.e., the ordering $\sigma_1 > \sigma_2$), thus we conclude $V = I_2$. We have established that (153) holds with equality iff $\boldsymbol{G} = U\Sigma$. This condition is equivalent to the condition in the statement. Specifically, the forward direction is $O_T = \boldsymbol{G}^\top\boldsymbol{G} = \Sigma^2$. The backward direction is: if $O_T = D$ for some diagonal $D \succ 0$, then $\boldsymbol{G}^\top\boldsymbol{G} = D$ which implies there exists some orthonormal $U \in \mathbb{R}^{T\times d}$ such that $\boldsymbol{G} = UD^{1/2}$, so D is a diagonal matrix of the squared singular values of $\boldsymbol{G}$. □

Lemma S.2. Let $z \in \mathbb{N}_0^n$ where $z_i \sim \text{Poi}(\lambda_i)$ is an independent count for some rate $\lambda_i > 0$. Let $x = \sum_{i=1}^n z_i \in \mathbb{N}_0$. Then $p(z|x) = \text{Mult}(x, \bar{\lambda})(z)$ where $\bar{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}$.

Proof. This is a consequence of the fact that $x \sim \text{Poi}(\Lambda)$ where $\Lambda = \sum_{i=1}^n \lambda_i$. Then for any $z \in \mathbb{N}_0^n$ and $x \in \mathbb{N}_0$ such that $x = \sum_{i=1}^n z$,

$$p(z,x) = p(z) = \prod_{i=1}^n \frac{\lambda_i^{z_i} e^{-\lambda_i}}{z_i!} \qquad\qquad p(x) = \frac{\Lambda^x e^{-\Lambda}}{x!}$$

so that

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{\left(\prod_i \lambda_i^{z_i}\right)\left(e^{-\Lambda}\right)}{\prod_i z_i!}\frac{x!}{\Lambda^x e^{-\Lambda}} = \frac{x!}{\prod_i z_i!}\frac{\prod_i \lambda_i^{z_i}}{\Lambda^x} = \frac{x!}{\prod_i z_i!}\frac{\left(\prod_i \bar{\lambda}_i^{z_i}\right)\left(\Lambda^x\right)}{\Lambda^x} = \frac{x!}{\prod_i z_i!}\prod_i \bar{\lambda}_i^{z_i} = \text{Mult}(x,\bar{\lambda})(z)$$

□

Lemma S.3. Heavy-ball SGD (25–26) with constant learning rate can achieve the following regret bound:

$$\sum_{t=1}^T l_t(w_t) - l_t(u) \le D_1 L\sqrt{T}\sqrt{\frac{1+\mu}{1-\mu}} \tag{155}$$

where $L \ge \max_t ||g_t||_2$ is the bound on the gradient norm. We additionally assume that l_t is L-Lipschitz.

Proof. We use the following claims without proof which is tedious.

Claim I. Define $\beta = \frac{\mu}{1-\mu}$ and $y_t = w_t + \beta(w_t - w_{t-1})$. Then $y_{t+1} = y_t - \frac{\eta}{1-\mu}g_t$.

Claim II. $||v_t||_2 \le \frac{L}{1-\mu}$ for all t.

From (20) and Claim I, we have

$$\sum_{t=1}^T l_t(y_t) - l_t(u) \le \frac{D_1^2(1-\mu)}{2\eta} + \frac{\eta T L^2}{2(1-\mu)} \tag{156}$$

Since l_t is L-Lipschitz, from Claim II we have

$$|l_t(w_t) - l_t(y_t)| \le L\,||w_t - y_t||_2 = L\beta\eta\,||v_{t-1}||_2 \le \frac{\eta L^2\mu}{(1-\mu)^2}$$

(the equality follows from $w_t - y_t = -\beta(w_t - w_{t-1}) = \beta\eta v_{t-1}$). Thus

$$\sum_{t=1}^T l_t(w_t) - l_t(y_t) \le \frac{\eta T L^2\mu}{(1-\mu)^2} \tag{157}$$

Adding (156) and (157), we get

$$\sum_{t=1}^T l_t(w_t) - l_t(u) \le \frac{D_1^2(1-\mu)}{2\eta} + \frac{\eta T L^2}{2(1-\mu)}\left(1 + \frac{2\mu}{1-\mu}\right)$$

Choosing $\eta = \frac{D_1}{L}\sqrt{\frac{(1-\mu)^3}{(1+\mu)T}}$ yields (155) □

Lemma S.4. For any twice-differentiable $f : \mathbb{R} \to \mathbb{R}$,

$$f(1) = f(0) + f'(0) + \int_0^1 (1-t) f''(t) dt$$

Proof. Applying the FTC twice gives us

$$\begin{aligned}
f(1) &= f(0) + \int_0^1 f'(s) ds \\
&= f(0) + \int_0^1 \left(f'(0) + \int_0^s f''(t) dt \right) ds \\
&= f(0) + f'(0) + \int_0^1 \int_0^s f''(t) dt \, ds \\
&= f(0) + f'(0) + \int_0^1 \int_t^1 f''(t) ds \, dt \qquad \text{(switching integral order for } 0 \le s \le t \le 1) \\
&= f(0) + f'(0) + \int_0^1 (1-t) f''(s) dt
\end{aligned}$$

□

Lemma S.5. Assume $J : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth (i.e., $J(v) - J(u) \le \nabla J(u)^\top (v-u) + \frac{L}{2} ||v-u||_2^2$). Assume we perform T SGD steps:

$$g_t = \nabla J(w_t) \qquad\qquad w_{t+1} = w_t - \eta g_t$$

Let $w^\star = \arg\min_{w \in \mathbb{R}^d} J(w)$ denote a minimizer. Choosing $\eta = \frac{1}{L}$ gives us the following convergence rate:

$$J(w_{T+1}) - J^\star \le \frac{L \, ||w_1 - w^\star||_2^2}{2T} \tag{158}$$

Proof. From the smoothness of J

$$J(w_t) - J(w_{t+1}) \ge \frac{||g_t||_2^2}{2L} \ge 0 \tag{159}$$

Namely, an update never increases the loss. Using the convexity of J as usual, we have

$$J(w_t) - J^\star \le g_t^\top (w_t - w^\star) \tag{160}$$

Also as usual, we express $g_t^\top (w_t - w^\star)$ as the difference between $||w_t - w^\star||_2^2$ and $||w_{t+1} - w^\star||_2^2$ by the SGD update:

$$\begin{aligned}
||w_{t+1} - w^\star||_2^2 &= \left|\left| w_t - w^\star - \frac{1}{L} g_t \right|\right|_2^2 = ||w_t - w^\star||_2^2 + \frac{1}{L^2} ||g_t||_2^2 - \frac{2}{L} g_t^\top (w_t - w^\star) \\
\Rightarrow \quad g_t^\top (w_t - w^\star) &= \frac{L}{2} \left(||w_t - w^\star||_2^2 - ||w_{t+1} - w^\star||_2^2 \right) + \frac{||g_t||_2^2}{2L} \\
&\le \frac{L}{2} \left(||w_t - w^\star||_2^2 - ||w_{t+1} - w^\star||_2^2 \right) + J(w_t) - J(w_{t+1})
\end{aligned}$$

The last inequality uses (159). Plugging this in (160), we have

$$J(w_{t+1}) - J^\star \le \frac{L}{2} \left(||w_t - w^\star||_2^2 - ||w_{t+1} - w^\star||_2^2 \right)$$

The RHS telescopes when summed over $t = 1 \dots T$, giving us

$$\sum_{t=1}^{T} J(w_{t+1}) - J^\star \le \frac{L \, ||w_1 - w^\star||_2^2}{2}$$

Since $J(w_{T+1}) - J^\star \le J(w_t) - J^\star$ for all t by (159), it implies (158). □

Lemma S.6. Assume $J : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex. Let $w^\star = \arg\min_{w\in\mathbb{R}^d} J(w)$ denote the unique minimizer. Define the **Newton ball** as $\mathcal{B} = \left\{w : ||w - w^\star||_2 \le \frac{m}{L}\right\}$. Assume that the Hessian is L-Lipschitz in the Newton ball in operator norm for $||\cdot||_2$ (i.e., $\left|\left|\nabla^2 J(u) - \nabla^2 J(v)\right|\right|_2 \le L\,||u - v||_2$ for $u, v \in \mathcal{B}$). Assume we perform the Newton steps:

$$g_t = \nabla J(w_t) \qquad\qquad w_{t+1} = w_t - H_t^{-1} g_t$$
$$H_t = \nabla^2 J(w_t)$$

If $w_t \in \mathcal{B}$, we have

$$||w_{t+1} - w^\star||_2 \le \frac{L}{2m}\,||w_t - w^\star||_2^2 \tag{161}$$

Proof. Let $\bar{H}_t = \int_0^1 \nabla^2 J(w_t + s(w^\star - w_t))ds$ and use the exact mean-Hessian gradient expression $g_t = \bar{H}_t(w_t - w^\star)$ (118) to write

$$\begin{aligned} w_{t+1} - w^\star &= w_t - w^\star - H_t^{-1} g_t \\ &= w_t - w^\star - H_t^{-1}\bar{H}_t(w_t - w^\star) \\ &= (I_d - H_t^{-1}\bar{H}_t)(w_t - w^\star) \end{aligned}$$

Thus

$$\begin{aligned} ||w_{t+1} - w^\star||_2 &\le \left|\left|I_d - H_t^{-1}\bar{H}_t\right|\right|_2 ||w_t - w^\star||_2 \\ &= \left|\left|H_t^{-1}(H_t - \bar{H}_t)\right|\right|_2 ||w_t - w^\star||_2 \\ &\le \left|\left|H_t^{-1}\right|\right|_2 \left|\left|\bar{H}_t - H_t\right|\right|_2 ||w_t - w^\star||_2 \\ &\le \frac{L}{2m}\,||w_t - w^\star||_2^2 \qquad \text{(strong convexity, Lipschitz Hessian)} \end{aligned}$$

To see the last inequality, from strong convexity we have $H_t \succeq mI_d$ (Lemma J.1) which implies $H_t^{-1} \preceq (1/m)I_d$. Taking the spectral norm on both sides gives $\left|\left|H_t^{-1}\right|\right|_2 \le 1/m$. Also

$$\begin{aligned} \left|\left|\bar{H}_t - H_t\right|\right|_2 &= \left|\left|\int_0^1 \left(\nabla^2 J(w_t + s(w^\star - w_t)) - \nabla^2 J(w_t)\right) ds\right|\right|_2 \\ &\le \int_0^1 \left|\left|\nabla^2 J(w_t + s(w^\star - w_t)) - \nabla^2 J(w_t)\right|\right|_2 ds \qquad (162) \\ &\le L\int_0^1 s\,||w^\star - w_t||\,ds \qquad \text{(Hessian is } L\text{-Lipschitz near } w^\star) \\ &= \frac{L}{2}\,||w^\star - w_t|| \end{aligned}$$

where (162) follows because for any matrix function $A : \mathbb{R} \to \mathbb{R}^{d\times d}$ and unit vector $v \in \mathbb{R}^d$

$$\left|\left|\left(\int_0^1 A(s)ds\right)v\right|\right|_2 = \left|\left|\int_0^1 A(s)vds\right|\right|_2 \overset{\text{triangle ineq.}}{\le} \int_0^1 ||A(s)v||_2\,ds \le \int_0^1 ||A(s)||_2\,ds$$

From the definition $||B||_2 := \sup_{v:||v||_2=1} ||Bv||_2$, it follows that $\left|\left|\int_0^1 A(s)ds\right|\right|_2 \le \int_0^1 ||A(s)||_2\,ds$. □

Corollary S.7. Assume the setting in Lemma S.6. If $w_t \in \mathcal{B}$, then for all $k \ge 1$

$$||w_{t+k} - w^\star||_2 \le \frac{1}{2^k}\,||w_t - w^\star||_2 \le \frac{m}{L}2^{-k} \tag{163}$$

In particular, $w_{t+k} \in \mathcal{B}$.

Proof.

$$||w_{t+1} - w^\star||_2 \le \left(\frac{L}{2m}\,||w_t - w^\star||_2\right)||w_t - w^\star||_2 \le \frac{1}{2}\,||w_t - w^\star||_2$$

The first inequality is (161). The second inequality holds since $w_t \in \mathcal{B}$ (i.e., $||w_t - w^\star|| \le \frac{m}{L}$). (163) follows by induction. □

Lemma S.8. Assume the setting in Lemma S.6. Assume further that $J(w)$ is M-smooth. Let $E_t = J(w_t) - J^\star \geq 0$ and $C = \frac{ML^2}{2m^4} > 0$. If $w_t \in \mathcal{B}$, then for all $k \geq 1$

$$E_{t+k} \leq C^{2^k-1} E_t^{2^k} = \frac{(CE_t)^{2^k}}{C} \tag{164}$$

Proof.

$$\begin{aligned}
E_{t+1} &\leq \frac{M}{2} ||w_{t+1} - w^\star||_2^2 && \text{(smoothness)} \\
&\leq \frac{M}{2} \left(\frac{L}{2m} ||w_t - w^\star||_2^2 \right)^2 && (161) \\
&= \frac{ML^2}{8m^2} ||w_t - w^\star||_2^4 \\
&\leq \frac{ML^2}{8m^2} \left(\frac{2}{m} E_t \right)^2 && \text{(strong convexity)} \\
&= \frac{ML^2}{2m^4} E_t^2 && (165)
\end{aligned}$$

Thus $E_{t+1} \leq CE_t^2$. (164) follows by induction. □

Corollary S.9. Assume the setting in Lemma S.8. If $w_t \in \mathcal{B}$, then for all $k \geq 1$

$$E_{t+k} \leq \frac{Mm^2}{2L^2} 4^{-k} \tag{166}$$

In particular, for $k > \log_4(\frac{M^2}{2m^2})$ we must have $CE_{t+k} < \frac{1}{2}$.

Proof. This follows from the M-smoothness $E_{t+k} \leq \frac{M}{2} ||w_{t+k} - w^\star||_2^2$ combined with the fact that the distance to $w^\star$ halves (Corollary S.7). □

Lemma S.10. (28–31) is equivalent to

$$\begin{aligned}
g_t' &= \nabla l_t(x_t) \\
v_t &= g_t' + \mu_t v_{t-1} \\
x_{t+1} &= x_t - \eta_t(g_t' + \mu_t v_t) + (\eta_t \mu_t - \eta_{t+1}\mu_{t+1}) v_t
\end{aligned}$$

Proof. We have

$$\begin{aligned}
w_{t+1} &= w_t - \eta_t v_t \\
&= x_t + \eta_t \mu_t v_{t-1} - \eta_t v_t \\
&= x_t - \eta_t (v_t - \mu_t v_{t-1}) \\
&= x_t - \eta_t g_t' \qquad (167)
\end{aligned}$$

Combining (28) and (167), we have

$$\begin{aligned}
x_{t+1} &= w_{t+1} - \eta_{t+1}\mu_{t+1} v_t \\
&= x_t - \eta_t g_t' - \eta_t \mu_t v_t + \eta_t \mu_t v_t - \eta_{t+1}\mu_{t+1} v_t \\
&= x_t - \eta_t (g_t' + \mu_t v_t) + (\eta_t \mu_t - \eta_{t+1}\mu_{t+1}) v_t \qquad (168)
\end{aligned}$$

□

Lemma S.11. Let $X_t = \beta X_{t-1} + (1-\beta) Z_t$ where $0 \leq \beta < 1$, $X_0 = 0$, and $Z_1, Z_2, \ldots \sim \mathbf{Unk}(\mu, \sigma^2)$ are iid. Then

$$\mathbf{E}[X_t] = (1 - \beta^t)\mu \quad \Rightarrow \quad \lim_{t\to\infty} \mathbf{E}[X_t] = \mu \tag{169}$$

$$\text{Var}(X_t) = \sigma^2 \frac{1-\beta}{1+\beta}(1 - \beta^{2t}) \quad \Rightarrow \quad \lim_{t\to\infty} \text{Var}(X_t) = \sigma^2 \frac{1-\beta}{1+\beta} \tag{170}$$

Proof. We can easily verify the form

$$X_t = (1-\beta)\sum_{k=0}^{t-1}\beta^k Z_{t-k} = (1-\beta)\left(\beta^{t-1}Z_1 + \cdots + \beta Z_{t-1} + Z_t\right)$$

Using $\sum_{k=0}^{t-1}\beta^k = (1-\beta^t)/(1-\beta)$, it follows that

$$\mathbf{E}[X_t] = \mathbf{E}\left[(1-\beta)\sum_{k=0}^{t-1}\beta^k Z_{t-k}\right] = (1-\beta)\sum_{k=0}^{t-1}\beta^k \mathbf{E}\left[Z_{t-k}\right] = \mu(1-\beta)\left(\sum_{k=0}^{t-1}\beta^k\right) = (1-\beta^t)\mu$$

and

$$\begin{aligned}\mathrm{Var}\left(X_t\right) = (1-\beta)^2\mathrm{Var}\left(\sum_{k=0}^{t-1}\beta^k Z_{t-k}\right) = (1-\beta)^2\sum_{k=0}^{t-1}\mathrm{Var}\left(\beta^k Z_{t-k}\right) = (1-\beta)^2\sum_{k=0}^{t-1}\beta^{2k}\sigma^2 &= \sigma^2\frac{(1-\beta)^2}{1-\beta^2}(1-\beta^{2t}) \\ &= \sigma^2\frac{1-\beta}{1+\beta}(1-\beta^{2t})\end{aligned}$$

□

Lemma S.12. For any positive constant $a > 0$, $\sqrt{a+\Theta(\epsilon)} = \sqrt{a} + \Theta(\epsilon)$ where the asymptotic term is in $\epsilon \to 0^+$.

Proof. Let $r : \mathbb{R} \to \mathbb{R}$ denote the asymptotic term on the LHS. Since $r(\epsilon) = \Theta(\epsilon)$, for all small enough $0 < \epsilon \le \epsilon_0$ we have $c_1\epsilon \le |r(\epsilon)| \le c_2\epsilon$ for some constants $c_1, c_2 > 0$. It also imples there exists $0 < \epsilon_1 \le \epsilon_0$ such that $|r(\epsilon)| \le a/2$ for all $0 < \epsilon \le \epsilon_1$. The key tool is the exact identity

$$\sqrt{a+r(\epsilon)} - \sqrt{a} = \frac{r(\epsilon)}{\sqrt{a+r(\epsilon)} + \sqrt{a}}$$

For all $0 < \epsilon \le \epsilon_1$, the denominator is at least $\sqrt{a/2}$ (since $a + r(\epsilon) \ge a/2$) and at most $(1+\sqrt{3/2})\sqrt{a}$ (since $a + r(\epsilon) \le 3a/2$), which implies

$$\left(\frac{c_1}{(1+\sqrt{3/2})\sqrt{a}}\right)\epsilon \le \left|\sqrt{a+r(\epsilon)} - \sqrt{a}\right| \le \left(\frac{c_2}{\sqrt{a/2}}\right)\epsilon$$

Therefore, $\left|\sqrt{a+r(\epsilon)} - \sqrt{a}\right| = \Theta(\epsilon)$. □

Lemma S.13. For any positive constant $a > 0$, $\sqrt{a+O(\epsilon)} = \sqrt{a} + O(\epsilon)$ where the asymptotic term is in $\epsilon \to 0^+$.

Proof. The proof is similar to/easier than the proof of Lemma S.12 (which however does not imply the current claim), so we leave it as an exercise. □

Lemma S.14 (Technical result for Lemma S.16)**.** $\mathbf{E}[M^2\Delta] = O(1-\beta_2)$ as $\beta_2 \to 1^-$.

Proof. Since

$$M = \sum_{i=0}^{\infty} a_i g_i \qquad a_i = (1-\beta_1)\beta_1^i \in [0,1)$$

$$\Delta = \sum_{k=0}^{\infty} b_k(g_k^2 - \sigma^2) \qquad b_k = (1-\beta_2)\beta_2^k \in [0,1)$$

we have

$$M^2\Delta = \sum_{i,j,k} a_i a_j b_k g_i g_j (g_k^2 - \sigma^2)$$

Since g_i, g_j and g_k^2 are iid with mean 0 and σ^2, the expectation is zero for all terms except at $i = j = k$. This implies

$$\begin{aligned}\mathbf{E}[M^2\Delta] = \sum_{i=0}^{\infty}\mathbf{E}[a_i^2 b_i g_i^2(g_i^2-\sigma^2)] = (m_4-\sigma^4)\sum_{i=0}^{\infty}a_i^2 b_i = (m_4-\sigma^4)(1-\beta_1)^2(1-\beta_2)\sum_{i=0}^{\infty}\beta_1^{2i}\beta_2^i \\ = \frac{(m_4-\sigma^4)(1-\beta_1)^2(1-\beta_2)}{1-\beta_1^2\beta_2}\end{aligned}$$

where $m_4 = \mathbf{E}[g^4]$ is the 4th moment of the gradient. Ignoring terms that are constant in β_2, we have

$$\frac{1-\beta_2}{1-\beta_1^2\beta_2} = \frac{1-\beta_2}{1-\beta_1^2+\beta_1^2(1-\beta_2)} = O(1-\beta_2)$$

To see the last step, note that the denomiator converges to a positive constant as $\beta_2 \to 1^-$. □

Lemma S.15 (Technical result for Lemma S.16)**.** $\mathbf{E}[\Delta^4] = O((1-\beta_2)^2)$ as $\beta_2 \to 1^-$.

Proof. Recall

$$\Delta = \sum_{k=0}^{\infty} b_k(g_k^2 - \sigma^2) \qquad\qquad b_k = (1-\beta_2)\beta_2^k \in [0,1)$$

where $X_k := g_k^2 - \sigma^2$ are iid with $\mathbf{E}[X_k] = 0$, $\mathbf{E}[X_k^2] = \tau^2$, and $\mathbf{E}[X_k^4] = \kappa_4 < \infty$. Then

$$\mathbf{E}[\Delta^4] = \sum_{i,j,k,l} \mathbf{E}[b_i b_j b_k b_l X_i X_j X_k X_l] = (\kappa_4 - 3\tau^4)\sum_i b_i^4 + 3\tau^4\left(\sum_i b_i^2\right)^2$$

The last inequality is standard and holds from the fact that there are 3 pairings over indices $((i:j,k:l)$, $(i:k,j:l)$, $(i:l,j:k))$, each contributing $\tau^4\sum_{i\neq j} b_i^2 b_j^2$ summing to $3\tau^4((\sum_i b_i^2)^2 - \sum_i b_i^4)$. The second term dominates. Specifically,

$$\sum_i b_i^4 = (1-\beta_2)^4\sum_i \beta_2^{4i} = \frac{(1-\beta_2)^4}{1-\beta_2^4} = \frac{(1-\beta_2)^4}{(1-\beta_2)(1+\beta_2+\beta_2^2+\beta_2^3)} = O((1-\beta_2)^3)$$

(the denominator tends to $4(1-\beta_2)$ as $\beta_2 \to 1^-$), while

$$\sum_i b_i^2 = \frac{(1-\beta_2)^2}{1-\beta_2^2} = O(1-\beta_2) \qquad\Rightarrow\qquad \left(\sum_i b_i^2\right)^2 = O((1-\beta_2)^2)$$

where $(1-\beta_2)^3 < (1-\beta_2)^2$. Thus $\mathbf{E}[\Delta^4] = O((1-\beta_2)^2)$. □

Lemma S.16. Let $M \sim \mathbf{Mom}_1(0, \sigma^2\frac{1-\beta_1}{1+\beta_1})$ and $V \sim \mathbf{Mom}_2(\sigma^2, \tau^2\frac{1-\beta_2}{1+\beta_2})$ where $\beta_1, \beta_2 \in [0,1)$ and $\sigma^2, \tau^2 > 0$. Assume $\sup_{\beta_2\in[0,1)} \mathbf{E}[1/V^2] \le C$. Define $O = \frac{M}{\sqrt{V}}$. Then

$$\sqrt{\mathbf{E}[O^2]} = \sqrt{\frac{1-\beta_1}{1+\beta_1}} + O(1-\beta_2) \tag{171}$$

where the asymptotic term is in $\beta_2 \to 1^-$.

Proof. Recall $\mathbf{E}[O^2] = \mathbf{E}[M^2/V]$ yields the first term of (171) if V is treated as constant σ^2. More generally, we take a Taylor expansion $1/V = 1/\sigma^2 + \cdots$ and analyze the remainder terms. Let $\Delta = V - \sigma^2$. Let G denote the good event that Δ is small and B the bad event (i.e., its complement). We analyze

$$\mathbf{E}\left[O^2\right] = \mathbf{E}\left[\mathbb{1}_G O^2\right] + \mathbf{E}\left[\mathbb{1}_B O^2\right] \tag{172}$$

Good term. Specifically, if G is defined as the event $|\Delta| \le \sigma^2/2$, we can use the following equality

$$\frac{1}{1+\Delta/\sigma^2} = \sum_{k=0}^{\infty}\left(\frac{-\Delta}{\sigma^2}\right)^k =: \psi(\Delta) \le 2$$

to explicitly bound the Taylor expansion $1/V = 1/\sigma^2 - \Delta/\sigma^4 + \Delta^2/\sigma^6 - \cdots$ as

$$\frac{1}{V} = \frac{1}{\sigma^2+\Delta} = \frac{1}{\sigma^2}\frac{1}{1+\Delta/\sigma^2} = \frac{1}{\sigma^2}\psi(\Delta) = \frac{1}{\sigma^2} - \frac{\Delta}{\sigma^4} + \frac{\Delta^2}{\sigma^6}\psi(\Delta)$$

Thus the good term in (172) becomes

$$\mathbf{E}\left[\mathbb{1}_G O^2\right] = \frac{\mathbf{E}[\mathbb{1}_G M^2]}{\sigma^2} - \frac{\mathbf{E}\left[\mathbb{1}_G M^2\Delta\right]}{\sigma^4} + \frac{\mathbf{E}\left[\mathbb{1}_G \psi(\Delta) M^2\Delta^2\right]}{\sigma^6} \tag{173}$$

In the first term, we have $\mathbf{E}[\mathbb{1}_G M^2] = \mathbf{E}[M^2] - \mathbf{E}[\mathbb{1}_B M^2]$ where $\mathbf{E}[M^2] = \mathrm{Var}(M) = \sigma^2(1-\beta_1)/(1+\beta_1)$ and, by Cauchy-Schwarz, $\mathbf{E}[\mathbb{1}_B M^2] \le \sqrt{\Pr(B)}\sqrt{\mathbf{E}[M^4]}$. To bound the probability of the bad event, we use the 4th-moment Markov inequality

$$\Pr(B) = \Pr\left(|\Delta| > \frac{\sigma^2}{2}\right) \le \frac{E[\Delta^4]}{(\frac{\sigma^2}{2})^4} = O((1-\beta_2)^2)$$

where $\mathbf{E}[\Delta^4] = O((1-\beta_2)^2)$ can be shown by moment analysis (Lemma S.15). The use of the 4th-moment gives a sharper bound than Chebyshev (i.e., 2nd-moment Markov), which yields $\Pr(B) = O(1-\beta_2)$ that is too loose to be useful. Combining together, we have shown that the first term of (173) is

$$\frac{\mathbf{E}[\mathbb{1}_G M^2]}{\sigma^2} = \frac{1-\beta_1}{1+\beta_1} + O(1-\beta_2)$$

We now move on to show that the other terms are $O(1-\beta_2)$. Since the presence of the indicator variable only makes the bound tighter (i.e., $|\mathbf{E}[\mathbb{1}_G X]| \le |\mathbf{E}[X]|$), we ignore it in the following. The second term $\mathbf{E}[M^2\Delta]$ can be directly shown as $O(1-\beta_2)$ (Lemma S.14). For the third term, we use the boundedness of $\psi(\Delta) \le 2$ and again the strong 4th-moment bound $\mathbf{E}[\Delta^4] = O((1-\beta_2)^2)$ to have $\left|\mathbf{E}\left[\psi(\Delta)M^2\Delta^2\right]\right| \le 2\left|\mathbf{E}\left[M^2\Delta^2\right]\right| \le 2\sqrt{\mathbf{E}[M^4]}\sqrt{\mathbf{E}[\Delta^4]} = O(1-\beta_2)$.

Bad term. We have

$$\left|\mathbf{E}\left[\mathbb{1}_B O^2\right]\right| \le \sqrt{\Pr(B)}\sqrt{\mathbf{E}[O^4]} = O(1-\beta_2)\sqrt{\mathbf{E}\left[\frac{M^4}{V^2}\right]} = O(1-\beta_2)$$

Here, we do not have an explicit control over the behavior of $1/V^2$ (e.g., it may assign too much mass near zero), thus we invoke the regularity assumption $\mathbf{E}[1/V^2] = O(1)$.

Together. Going back to (172), we have

$$\mathbf{E}\left[O^2\right] = \mathbf{E}\left[\mathbb{1}_G O^2\right] + \mathbf{E}\left[\mathbb{1}_B O^2\right] = \left(\frac{1-\beta_1}{1+\beta_1} + O(1-\beta_2)\right) + O(1-\beta_2) = \frac{1-\beta_1}{1+\beta_1} + O(1-\beta_2)$$

Finally, it follows from Lemma S.13 that $\sqrt{\mathbf{E}[O^2]} = \sqrt{\frac{1-\beta_1}{1+\beta_1}} + O(1-\beta_2)$. □

Lemma S.17. (132) ⇔ (133) ⇒ (134)

Proof. (132) ⇒ (133): To exploit the Hessian bound (132) for a gradient difference, we can use the Jacobian-based FTC (131).

$$||\nabla f(x) - \nabla f(z)||_* = \left|\left|\int_0^1 \nabla^2 f(z + r(x-z))(x-z)dr\right|\right|_* \le \int_0^1 \left|\left|\nabla^2 f(z + r(x-z))(x-z)\right|\right|_* dr \le L\,||x-z||$$

(133) ⇒ (132): Pick any $h \in \mathbb{R}^d$. To exploit the gradient difference bound (133) for a Hessian, we can use the Taylor expansion of the gradient $\nabla f(x+rh) \approx \nabla f(x) + r\nabla^2 f(x)h$. For any $r \ne 0$,

$$||\nabla f(x+rh) - \nabla f(x)||_* \le L\,|r|\,||h|| \qquad \Rightarrow \qquad \left|\left|\frac{\nabla f(x+rh) - \nabla f(x)}{r}\right|\right|_* \le L\,||h||$$

Taking the limit $r \to 0$ gives $\left|\left|\nabla^2 f(x)h\right|\right|_* \le L\,||h||$.
(133) ⇒ (134): To exploit the gradient difference bound (133) for an output difference, we can use the gradient-based FTC (130).

$$\begin{aligned}
|f(x) - f(z) - \langle \nabla f(z), x-z\rangle| &= \left|\int_0^1 \langle \nabla f(z + r(x-z)) - \nabla f(z), x-z\rangle\, dr\right| \\
&\le \int_0^1 |\langle \nabla f(z + r(x-z)) - \nabla f(z), x-z\rangle|\, dr \\
&\le \int_0^1 ||\nabla f(z + r(x-z)) - \nabla f(z)||_*\, ||x-z||\, dr \\
&\le \int_0^1 (Lr\,||x-z||)\, ||x-z||\, dr \\
&= \frac{L}{2}\,||x-z||^2
\end{aligned}$$

□

Lemma S.18. When $||\cdot|| = ||\cdot||_2$, we have: (132) $\Leftrightarrow$ (133) $\Leftrightarrow$ (134)

Proof. Since (132) $\Leftrightarrow$ (133) $\Rightarrow$ (134) always (Lemma S.17), it is sufficient to show (134) $\Rightarrow$ (132). For the Euclidean norm, we only need to show $\left|\left|\nabla^2 f(x)\right|\right|_2 \leq L$. To exploit the linearization error bound (134) for a Hessian, we note that the Hessian is already the main term in the linearization error. Pick any point $x \in \mathbb{R}^d$ and direction $u \in \mathbb{R}^d$. For any $r \neq 0$,

$$\left|f(x+ru) - f(x) - r\nabla f(x)^\top u\right| = \left|\frac{r^2}{2} u^\top \nabla^2 f(x) u + o(r^2)\right| \leq \frac{Lr^2}{2} ||u||^2$$

Dividing by $r^2/2$ and taking the limit $r \to 0$,

$$\lim_{r\to 0} \left|u^\top \nabla^2 f(x) u + \frac{o(r^2)}{r^2}\right| = \left|u^\top \nabla^2 f(x) u\right| \leq L ||u||^2$$

The last inequality is equivalent to $\left|\left|\nabla^2 f(x)\right|\right|_2 \leq L$. □

Lemma S.19. Let $w' = (r/||w - \eta g||_2)(w - \eta g)$ for nonzero $w, g \in \mathbb{R}^d$ with $||w||_2 = r > 0$ and $\eta > 0$. Then

$$w' = w - \eta g^\perp + O(\eta^2)$$

where $g^\perp \in \mathbb{R}^d$ is the projection of $g \in \mathbb{R}^d$ onto $\text{span}\,(w)^\perp$.

Proof. Since $||w - \eta g||_2^2 = r^2 - 2\eta w^\top g + \eta^2 ||g||_2^2$, we can write the renormalization as

$$\frac{r}{||w - \eta g||_2} = \left(r - 2\frac{\eta w^\top g}{r^2} + O(\eta^2)\right)^{-1/2} = 1 + \frac{\eta w^\top g}{r^2} + O(\eta^2)$$

where the last equality uses $(1+x)^{-1/2} = 1 - x/2 + O(x^2)$ with $x = -2\eta w^\top g/r^2 + O(\eta^2)$. Plugging it in the update, we have

$$w' = w - \eta\left(I_{d\times d} - \frac{ww^\top}{r^2}\right) g + O(\eta^2)$$

□

Lemma S.20. Let $\text{Sym}(d)$ denote $d \times d$ symmetric matrices. Given any square matrix $M \in \mathbb{R}^{d\times d}$, let

$$\text{sym}(M) := \frac{M + M^\top}{2} \qquad\qquad \text{skew}(M) := \frac{M - M^\top}{2}$$

Note $M = \text{sym}(M) + \text{skew}(M)$; $M \in \text{Sym}(d)$ iff $M = \text{sym}(M)$; and $\text{skew}(M) = -\text{skew}(M)^\top$. Then

$$\langle M, S\rangle = \langle \text{sym}(M), S\rangle \qquad\qquad \forall S \in \text{Sym}(d)$$

Proof. It is sufficient to show $\langle \text{skew}(M), S\rangle = 0$. For this we note

$$\langle \text{skew}(M), S\rangle = \text{tr}\left(\text{skew}(M)^\top S\right) = \text{tr}\left(S\text{skew}(M)\right) = -\text{tr}\left(S\text{skew}(M)^\top\right) = -\text{tr}\left(\text{skew}(M)^\top S\right)$$

Since $\text{tr}\left(\text{skew}(M)^\top S\right) = -\text{tr}\left(\text{skew}(M)^\top S\right)$, we must have $\text{tr}\left(\text{skew}(M)^\top S\right) = 0$. □

Lemma S.21. Given nonzero $w, g \in \mathbb{R}^d$ and $\eta > 0$, the maximizer of

$$f^\star = \max_{\Delta\in\mathbb{R}^d:\ ||\Delta||\leq\eta,\ \Delta^\top w = 0} \Delta^\top g \tag{174}$$

is given by $\Delta^\star = \eta z$ for any $z \in \mathbb{R}^d$ satisfying $z \in \partial\,||g + \lambda^\star w||_*$ and $w^\top z = 0$ where $\lambda^\star \in \arg\min_{\lambda\in\mathbb{R}} ||g + \lambda w||_*$. This z exists.

Proof. Note that for Δ already satisfying $||\Delta|| \leq \eta$, the Lagrangian can be defined as $L(\Delta, \lambda) = \Delta^\top (g + \lambda w)$. This implies that the dual function can be expressed in dual norm as follows:

$$g(\lambda) = \max_{\Delta \in \mathbb{R}^d:\, ||\Delta|| \leq \eta} \Delta^\top (g + \lambda w) = \eta \, ||g + \lambda w||_*$$

Since (174) satisfies Slater's condition, strong duality holds and $f^\star = \min_\lambda g(\lambda) = g(\lambda^\star) = \eta \, ||h^\star||_*$ where we denote $h^\star = g + \lambda^\star w$. So we just need to characterize a feasible Δ that achieves $\langle \Delta, g \rangle = \eta \, ||h^\star||_*$. We invoke the standard characterization of the subgradient of a norm (152):

$$z \in \partial \, ||h^\star||_* \qquad \Leftrightarrow \qquad ||z|| \leq 1 \wedge \langle z, h^\star \rangle = ||h^\star||_*$$

This implies $\Delta^\star = \eta z$ for $z \in \partial \, ||g + \lambda^\star w||_*$ orthogonal to w achieves $f^\star$. The last piece is showing the existence of such z. Let $\phi(\lambda) = ||g + \lambda w||_*$. Since $\lambda^\star$ minimizes ϕ we must have $0 \in \partial \phi(\lambda^\star)$. By the subgradient chain rule (150): $\partial \phi(\lambda^\star) = \{w^\top z : z \in \partial \, ||g + \lambda^\star w||_*\}$. So there exists $z \in \partial \, ||g + \lambda^\star w||_*$ such that $w^\top z = 0$. □

Lemma S.22. Let $W, G \in \mathbb{R}^{D \times d}$ with $D \times d$ and $\eta > 0$. The maximizer of

$$f^\star = \max_{\Delta \in \mathbb{R}^{D \times d}:\, ||\Delta|| \leq \eta,\, \Delta^\top W + W^\top \Delta = 0_{d \times d}} \langle \Delta, G \rangle \tag{175}$$

is given by $\Delta^\star = \eta Z$ for any $Z \in \mathbb{R}^{D \times d}$ satisfying $Z \in \partial \, ||G + W\Lambda^\star||_*$ and $Z^\top W + W^\top Z = 0_{d \times d}$ where $\Lambda^\star \in \arg\min_{\Lambda \in \mathrm{Sym}(d)} ||G + W\Lambda||_*$. This Z exists.

Proof. Note that for Δ already satisfying $||\Delta|| \leq \eta$, the Lagrangian can be defined as $L(\Delta, \Lambda) = \langle \Delta, G \rangle + (1/2) \left\langle \Lambda, \Delta^\top W + W^\top \Delta \right\rangle$. Since $\Delta^\top W + W^\top \Delta$ is symmetric, we can assume Λ is symmetric WLOG (Lemma S.20). In that case, we can write $(1/2) \left\langle \Lambda, \Delta^\top W + W^\top \Delta \right\rangle = \langle \Delta, W\Lambda \rangle$, yielding the dual function as dual norm:

$$g(\Lambda) = \max_{\Delta \in \mathbb{R}^d:\, ||\Delta|| \leq \eta} \langle \Delta, G + W\Lambda \rangle = \eta \, ||G + W\Lambda||_*$$

Since (175) satisfies Slater's condition, strong duality holds and $f^\star = \min_\Lambda g(\Lambda) = g(\Lambda^\star) = \eta \, ||H^\star||_*$ where we denote $H^\star = G + W\Lambda^\star$. So we just need to characterize a feasible Δ that achieves $\langle \Delta, G \rangle = \eta \, ||H^\star||_*$. We invoke the standard characterization of the subgradient of a norm (152):

$$Z \in \partial \, ||H^\star||_* \qquad \Leftrightarrow \qquad ||Z|| \leq 1 \wedge \langle Z, H^\star \rangle = ||H^\star||_*$$

This implies $\Delta^\star = \eta Z$ for $Z \in \partial \, ||G + W\Lambda^\star||_*$ satisfying $Z^\top W + W^\top Z = 0_{d \times d}$ achieves $f^\star$. The last piece is showing the existence of such Z. Let $\tilde{\phi} : \mathbb{R}^{d \times d} \to \mathbb{R}$ denote the unrestricted function $\tilde{\phi}(\Lambda) = ||G + W\Lambda||_*$. Since $\Lambda^\star$ minimizes $\tilde{\phi}$ over $\mathrm{Sym}(d)$, the convex optimality condition gives $0_{d \times d} \in \partial \tilde{\phi}(\Lambda^\star) + \mathrm{Sym}(d)^\perp$.[32] Thus $Y^\star + N^\star = 0_{d \times d}$ for some $Y^\star \in \partial \tilde{\phi}(\Lambda^\star)$ and $N^\star \in \mathrm{Sym}(d)^\perp$, but then $Y^\star = -N^\star$ is skew-only. At the same time, $Y^\star = W^\top Z$ for some $Z \in \partial \, ||G + W\Lambda^\star||_*$ by (151). Since $\mathrm{sym}(W^\top Z) = 0_{d \times d}$, we have $Z^\top W + W^\top Z = 0_{d \times d}$. □

[32] This follows from the fact that

$$\min_{\Lambda \in \mathrm{Sym}(d)} \tilde{\phi}(\Lambda) = \min_{\Lambda \in \mathbb{R}^{d \times d}} \left(\tilde{\phi}(\Lambda) + I_{\mathrm{Sym}(d)}(\Lambda) \right)$$

where $I_{\mathrm{Sym}(d)} : \mathbb{R}^{d \times d} \to \{0, \infty\}$ is the indicator of $\mathrm{Sym}(d)$. Thus at a convex minimizer $\Lambda^\star$, $0_{d \times d} \in \partial \tilde{\phi}(\Lambda^\star) + \partial I_{\mathrm{Sym}(d)}(\Lambda^\star)$. Finally, $\partial I_{\mathrm{Sym}(d)}(\Lambda^\star) = \mathrm{Sym}(d)^\perp$ (i.e., the indicator contributes nothing for symmetric directions).