Lecture 10: Adaptive Optimization and Well-structured Preconditioners

Lecture 9 analyzed mirror descent with one fixed mirror geometry for the whole run. In practice, however, gradient magnitudes, active directions, and local curvature are revealed only through the iterates. Lecture 10 studies adaptive first-order methods that choose a changing quadratic geometry from past gradients, with AdaGrad- and Shampoo-type methods as the guiding examples and a clean convex online-learning template as the analysis model.

Let E be a finite-dimensional real Hilbert space with inner product $\langle \cdot, \cdot \rangle_E$ and induced norm $\|\cdot\|_E$. Throughout the abstract analysis spine of this lecture, we use the finite-dimensional Riesz identification $E \simeq E^*$, so gradients and subgradients are regarded as vectors in E. Write $\mathcal{S}_{++}(E)$ and $\mathcal{S}_{+}(E)$ for the cones of self-adjoint positive-definite and self-adjoint positive-semidefinite operators on E, respectively, and write $\mathcal{L}(E)$ for the space of all linear operators on E.

For any $A \in \mathcal{S}_{+}(E)$ and $z \in E$, define the A-(semi)norm induced by the ambient inner product $\langle \cdot, \cdot \rangle_E$ as

$$\|z\|_A := \sqrt{\langle z, Az \rangle_E}.$$

This is a genuine norm when $A \succ 0$ and a seminorm when A is merely positive semidefinite. For $H \in \mathcal{S}_{++}(E)$, the dual norm of $\|\cdot\|_H$ (relative to $\langle \cdot, \cdot \rangle_E$) is $\|\cdot\|_{H^{-1}}$. In particular, $\|x\|_{I_E} = \|x\|_E$ so every H-norm is built on top of the ambient E-norm by reweighting via the operator H.

10.1 Motivation: From a Fixed Geometry to an Adaptive Target

The starting point is the fixed-geometry guarantee from Lecture 9. If we choose one quadratic mirror map and keep it for the whole run, then the usual mirror-descent telescope already tells us exactly how the chosen geometry enters the regret bound.

Definition 10.1 (Quadratic preconditioner geometry)**.** For a fixed operator $H \in \mathcal{S}_{++}(E)$, define

$$\Phi_H(x) := \frac{1}{2} \langle x, Hx \rangle_E .$$

Then

$$D_{\Phi_H}(x, y) = \frac{1}{2} \|x - y\|_H^2 .$$

Thus the mirror step generated by Φ_H with stepsize 1 is

$$x_{t+1} \in \arg\min_{x \in X} \Big\{ \langle g_t, x \rangle_E + \frac{1}{2} \|x - x_t\|_H^2 \Big\}.$$

So choosing a quadratic mirror map is exactly the same as choosing a positive-definite preconditioner H.

Corollary 10.1 (Fixed quadratic geometry)**.** *Let $X \subseteq E$ be nonempty, closed, and convex. Let $H \in \mathcal{S}_{++}(E)$, let $x_1, \dots, x_{T+1} \in X$, let $g_1, \dots, g_T \in E$, and suppose that*

$$x_{t+1} \in \arg\min_{x \in X} \left\{ \langle g_t, x \rangle_E + \frac{1}{2} \|x - x_t\|_H^2 \right\} \qquad \forall t \in \{1, \dots, T\}.$$

Then, for every $u \in X$,

$$\sum_{t=1}^{T} \langle g_t, x_t - u \rangle_E \le \frac{1}{2} \|u - x_1\|_H^2 \; + \; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 .$$

Proof of Corollary 10.1. Apply Corollary 9.4 to the quadratic mirror map

$$\Phi_H(x) := \frac{1}{2} \langle x, Hx \rangle_E ,$$

equipped with the norm $\|\cdot\|_H$, whose dual norm is $\|\cdot\|_{H^{-1}}$. Then Φ_H is 1-strongly convex with respect to $\|\cdot\|_H$, we have

$$D_{\Phi_H}(x, y) = \frac{1}{2} \|x - y\|_H^2 ,$$

and the displayed update is exactly the constant-stepsize mirror step with $\eta_t \equiv 1$. The conclusion follows from the constant-stepsize part of Corollary 9.4. □

For a feasible set X and a fixed metric H, following the structured-preconditioner paper, write

$$\|X\|_H := \sup_{x \in X} \|x\|_H .$$

Since $x_1, u \in X$, we have $\|u - x_1\|_H \le 2 \|X\|_H$, and Corollary 10.1 implies the uniform fixed-metric bound

$$\sum_{t=1}^{T} \langle g_t, x_t - u \rangle_E \le 2 \|X\|_H^2 \; + \; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 \qquad \forall u \in X.$$

The same metric H appears in two opposite ways: making H larger shrinks the inverse-gradient terms but inflates $\|X\|_H$; making H smaller does the opposite. If the full gradient sequence were known in advance, one could choose the best single H after the fact. For an admissible scale-invariant family $\mathcal{H}$, this fixed-geometry hindsight problem is

$$\min_{H \in \mathcal{H}} \left\{ 2 \|X\|_H^2 \; + \; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 \right\}.$$

The first simplification is to separate the scale of H from its shape. This is where the family-level version of the same radius enters.

Definition 10.2 (Family norm and domain radius)**.** Let $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ be nonempty and closed under positive scaling, and let $\bar{\mathcal{H}} \subseteq \mathcal{S}_+(E)$ denote its closure. For every $x \in E$, define

$$\|x\|_{\mathcal{H}} := \sup_{\substack{H \in \bar{\mathcal{H}} \\ \operatorname{tr}(H) \le 1}} \|x\|_H .$$

For every nonempty set $X \subseteq E$, define its possibly infinite $\mathcal{H}$-radius by

$$\|X\|_{\mathcal{H}} := \sup_{x \in X} \|x\|_{\mathcal{H}} .$$

The quantity $\|X\|_{\mathcal{H}}$ measures the size of X uniformly over normalized geometries generated by $\mathcal{H}$. Indeed, for every $H \in \mathcal{H}$, the normalized operator $H/\operatorname{tr}(H)$ still belongs to $\mathcal{H}$, and therefore

$$\|X\|_H^2 = \operatorname{tr}(H)\, \|X\|_{H/\operatorname{tr}(H)}^2 \le \operatorname{tr}(H)\, \|X\|_{\mathcal{H}}^2 .$$

Thus the fixed-geometry hindsight objective is controlled by

$$2\, \|X\|_{\mathcal{H}}^2 \operatorname{tr}(H) \;+\; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 .$$

Assume that $\|X\|_{\mathcal{H}} < \infty$; otherwise this upper bound is vacuous. Since $\mathcal{H}$ is closed under positive scaling, write $H = r\widehat{H}$, where $r = \operatorname{tr}(H) > 0$, $\widehat{H} := H/\operatorname{tr}(H) \in \mathcal{H}$, and $\operatorname{tr}(\widehat{H}) = 1$. For the scale-free, unregularized shape complexity of the gradient sequence, write

$$|||g_{1:T}|||_{\mathcal{H}} := \inf_{\substack{H \in \mathcal{H} \\ \operatorname{tr}(H)=1}} \sqrt{\sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2}.$$

Minimizing the scalar r first gives the product form

$$\begin{aligned} &\inf_{H \in \mathcal{H}} \left\{ 2\, \|X\|_{\mathcal{H}}^2 \operatorname{tr}(H) \;+\; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 \right\} \\ &\qquad = 2\, \|X\|_{\mathcal{H}}\; |||g_{1:T}|||_{\mathcal{H}}. \end{aligned}$$

This product is the clean offline comparator: twice the domain radius in the family norm times the best normalized inverse-metric gradient energy. An adaptive algorithm should try to achieve this quantity, up to constants, while seeing only past gradients. For designing such an algorithm, it is convenient not to optimize out the scalar r. Replacing the unknown domain scale by a tunable parameter η gives the additive offline surrogate

$$\min_{H \in \mathcal{H}} \left\{ \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 + \eta^2 \operatorname{tr}(H) \right\}.$$

This is still not an online rule, because it uses the full terminal gradient sequence. The rest of the lecture asks how to replace the terminal sum by observed prefixes and thereby choose H_t online.

This is not a scalar tuning problem. A diagonal preconditioner has d scale parameters and a full positive-definite geometry has $d(d+1)/2$, so external search is unrealistic in high dimension. The algorithm must learn the geometry from the observed gradients.

10.2 General Adaptive Metric Bounds

The offline surrogate suggests an adaptive metric. We first ask what can be proved for an arbitrary sequence $H_1, \dots, H_T$; the next subsection chooses this sequence online.

Definition 10.3 (Adaptive quadratic proxy update)**.** Let $X \subseteq E$ be nonempty, closed, and convex. For each $t \in \mathbb{N}$, let $x_t \in X$, let $g_t \in E$, and let $H_t \in \mathcal{S}_{++}(E)$. The adaptive quadratic proxy update is

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in X} \Big\{ \langle g_t, x \rangle_E + \frac{1}{2} \|x - x_t\|_{H_t}^2 \Big\}.$$

The one-step statement below is the fixed-metric inequality that will be summed with a changing choice of H_t.

Lemma 10.2 (One-step inequality for operator-preconditioned proxy minimization)**.** *Let $X \subseteq E$ be nonempty, closed, and convex. Let $H \in \mathcal{S}_{++}(E)$, let $x_t \in X$, let $g_t \in E$, and let*

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in X} \Big\{ \langle g_t, x \rangle_E + \frac{1}{2} \|x - x_t\|_H^2 \Big\}.$$

Then, for every $u \in X$,

$$\langle g_t, x_t - u \rangle_E \le \frac{1}{2} \|u - x_t\|_H^2 - \frac{1}{2} \|u - x_{t+1}\|_H^2 \; + \; \frac{1}{2} \|g_t\|_{H^{-1}}^2 .$$

Proof of Lemma 10.2. By the quadratic mirror-geometry remark above, the displayed update is exactly the constrained mirror step for the quadratic mirror map $\Phi_H(x) = \frac{1}{2} \langle x, Hx \rangle_E$. Therefore Theorem 8.9 with stepsize 1 gives, for every $u \in X$,

$$\langle g_t, x_t - u \rangle_E \le D_{\Phi_H}(u, x_t) - D_{\Phi_H}(u, x_{t+1}) \; + \; \langle g_t, x_t - x_{t+1} \rangle_E - D_{\Phi_H}(x_{t+1}, x_t).$$

Using $D_{\Phi_H}(x, y) = \frac{1}{2} \|x - y\|_H^2$, this becomes

$$\langle g_t, x_t - u \rangle_E \le \frac{1}{2} \|u - x_t\|_H^2 - \frac{1}{2} \|u - x_{t+1}\|_H^2 \; + \; \langle g_t, x_t - x_{t+1} \rangle_E - \frac{1}{2} \|x_{t+1} - x_t\|_H^2 .$$

Now Φ_H is 1-strongly convex with respect to the norm $\|\cdot\|_H$, whose dual norm is $\|\cdot\|_{H^{-1}}$. Hence Lemma 9.3 with $(x, y, g) = (x_t, x_{t+1}, g_t)$ yields

$$\langle g_t, x_t - x_{t+1} \rangle_E - \frac{1}{2} \|x_{t+1} - x_t\|_H^2 \le \frac{1}{2} \|g_t\|_{H^{-1}}^2 .$$

Substituting this bound into the previous display proves the claim. □

If the metrics are allowed to change arbitrarily, the quadratic distance terms need not telescope with a useful sign. The monotonicity condition $H_1 \preceq \cdots \preceq H_T$ is exactly what keeps the leftover terms positive and interpretable as metric growth.

Theorem 10.3 (General comparison bound under increasing metrics)**.** *Let $X \subseteq E$ be nonempty, closed, and convex. Let $H_1, \dots, H_T \in \mathcal{S}_{++}(E)$ satisfy*

$$H_1 \preceq H_2 \preceq \cdots \preceq H_T.$$

For each $t \in \{1, \dots, T\}$, let $x_t \in X$, let $g_t \in E$, and let

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in X} \Big\{ \langle g_t, x \rangle_E + \frac{1}{2} \|x - x_t\|_{H_t}^2 \Big\}.$$

Then, for every $u \in X$,

$$\sum_{t=1}^{T} \langle g_t, x_t - u \rangle_E \le \frac{1}{2} \|u - x_1\|_{H_1}^2 \; + \; \frac{1}{2} \sum_{t=2}^{T} \|u - x_t\|_{H_t - H_{t-1}}^2 \; + \; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 .$$

Moreover, if $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ is nonempty and closed under positive scaling, with closure $\bar{\mathcal{H}} \subseteq \mathcal{S}_+(E)$, and if

$$H_1 \in \bar{\mathcal{H}}, \qquad H_t - H_{t-1} \in \bar{\mathcal{H}} \quad \forall t \in \{2, \dots, T\},$$

then

$$\sum_{t=1}^{T} \langle g_t, x_t - u \rangle_E \le 2 \|X\|_{\mathcal{H}}^2 \operatorname{tr}(H_T) \; + \; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 .$$

Proof of Theorem 10.3. Apply Lemma 10.2 at each time t with the same comparator u. Summing over $t \in \{1, \dots, T\}$ gives

$$\begin{aligned}
\sum_{t=1}^{T} \langle g_t, x_t - u \rangle_E &\le \frac{1}{2} \sum_{t=1}^{T} \Big(\|u - x_t\|_{H_t}^2 - \|u - x_{t+1}\|_{H_t}^2 \Big) + \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 \\
&\le \frac{1}{2} \|u - x_1\|_{H_1}^2 + \frac{1}{2} \sum_{t=2}^{T} \Big(\|u - x_t\|_{H_t}^2 - \|u - x_t\|_{H_{t-1}}^2 \Big) + \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 \\
&= \frac{1}{2} \|u - x_1\|_{H_1}^2 + \frac{1}{2} \sum_{t=2}^{T} \|u - x_t\|_{H_t - H_{t-1}}^2 + \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 .
\end{aligned}$$

The inequality in the middle line drops the nonpositive terminal term $-\frac{1}{2} \|u - x_{T+1}\|_{H_T}^2$. This proves the first inequality.
Under the additional family assumption, the same argument is applied with the family norm. For $A \in \bar{\mathcal{H}} \cap \mathcal{S}_+(E)$, if $A \neq 0$, then

$$\|z\|_A^2 = \operatorname{tr}(A) \, \|z\|_{A/\operatorname{tr}(A)}^2 \le \operatorname{tr}(A) \, \|z\|_{\mathcal{H}}^2 ,$$

and the same bound is trivial if $A = 0$. Since $u, x_t \in X$, the triangle inequality gives $\|u - x_t\|_{\mathcal{H}} \le 2 \|X\|_{\mathcal{H}}$. Therefore

$$\|u - x_1\|_{H_1}^2 \le 4 \|X\|_{\mathcal{H}}^2 \operatorname{tr}(H_1), \qquad \|u - x_t\|_{H_t - H_{t-1}}^2 \le 4 \|X\|_{\mathcal{H}}^2 \operatorname{tr}(H_t - H_{t-1})$$

for $t \ge 2$. Substituting these estimates into the first inequality gives the strengthened family-radius bound. □

Theorem 10.3 reduces the adaptive problem to metric growth and inverse-metric gradient energy. The strengthened conclusion is the form needed by AdaReg: the selected metrics must not only be increasing, but their growth must occur inside the admissible family cone.

10.3 AdaReg: Choosing the Metrics Online

AdaReg turns the offline surrogate motivated by Corollary 10.1 into an online rule by replacing the unknown terminal second-moment operator by the currently available prefix. For $g \in E$, write

$$g \otimes_E g : E \to E, \qquad (g \otimes_E g)(x) := \langle g, x \rangle_E \, g.$$

This is the self-adjoint rank-one operator that records the direction of the observed gradient. Given $\varepsilon > 0$, define the cumulative second-moment operators

$$M_0 := \varepsilon I_E, \qquad M_t := M_{t-1} + g_t \otimes_E g_t = \varepsilon I_E + \sum_{s=1}^{t} (g_s \otimes_E g_s).$$

The regularization εI_E is not part of the hindsight motivation; it keeps every prefix M_t positive definite so that the selector is well defined from the first step.

The selector is a *be-the-leader* rule for the metric-design surrogate. After observing $g_{1:t}$, use the current-prefix hindsight metric

$$\min_{H \in \mathcal{H}} \Big\{ \operatorname{tr}(M_t H^{-1}) + \eta^2 \operatorname{tr}(H) \Big\}.$$

AdaReg uses this leader as the metric for the proxy step from x_t to x_{t+1}.

Algorithm 1 AdaReg meta-algorithm over a preconditioner family

Require: A nonempty closed convex set $X \subseteq E$, a family $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$, parameters $\eta > 0$, $\varepsilon > 0$, and an initial point $x_1 \in X$.

1: Set $M_0 \leftarrow \varepsilon I_E$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Receive a covector, identified with $g_t \in E$, at the current point x_t.
4: Update $M_t \leftarrow M_{t-1} + g_t \otimes_E g_t$.
5: Choose

$$H_t \in \arg\min_{H \in \mathcal{H}} \Big\{ \operatorname{tr}(M_t H^{-1}) + \eta^2 \operatorname{tr}(H) \Big\}.$$

6: Set

$$x_{t+1} \in \arg\min_{x \in X} \Big\{ \langle g_t, x \rangle_E + \frac{1}{2} \|x - x_t\|_{H_t}^2 \Big\}.$$

7: **end for**

Remark 10.1 (Why this selector objective?). The be-the-leader objective has the same two pieces as the fixed-metric tradeoff. The term

$$\operatorname{tr}(M_t H^{-1}) = \varepsilon \operatorname{tr}(H^{-1}) + \sum_{s=1}^{t} \|g_s\|_{H^{-1}}^2$$

penalizes choosing a metric H that is too small in directions where the historical gradients have been large. The competing term

$$\eta^2 \operatorname{tr}(H)$$

penalizes choosing the metric too large overall. Thus the selector is the prefix version of the fixed-metric tradeoff from the motivation section.

The selector estimate below deliberately keeps this raw objective, rather than optimizing out the scale of H. This is the quantity AdaReg is following online.

Here is the elementary online-learning fact behind the estimate. For fixed functions $\ell_0, \ell_1, \ldots, \ell_T$ on a common decision set, define the prefix objective

$$L_t(H) := \sum_{s=0}^{t} \ell_s(H).$$

If H_t is a minimizer of L_t, then H_t is the *leader* after seeing the first t losses. The "be-the-leader" inequality says that evaluating each loss ℓ_t at the leader that already knows ℓ_t is no worse than evaluating all losses at any fixed comparator H:

$$\sum_{t=0}^{T} \ell_t(H_t) \le \sum_{t=0}^{T} \ell_t(H).$$

AdaReg is allowed to use this current leader because g_t is observed before H_t is used in the quadratic proxy step. The next lemma applies this elementary fact to the metric-design losses.

Lemma 10.4 (Be-the-leader estimate for the metric selector)**.** *Let $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ be nonempty and closed under positive scaling, let $\eta > 0$, let $\varepsilon > 0$, and let $g_1, \dots, g_T \in E$. For each $t \in \{0, \dots, T\}$, suppose that*

$$H_t \in \operatorname*{arg\,min}_{H \in \mathcal{H}} \left\{ \varepsilon \operatorname{tr}(H^{-1}) \; + \; \sum_{s=1}^{t} \|g_s\|_{H^{-1}}^2 \; + \; \eta^2 \operatorname{tr}(H) \right\},$$

where the sum is empty for $t = 0$. Then

$$\sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 \le \inf_{H \in \mathcal{H}} \left\{ \varepsilon \operatorname{tr}(H^{-1}) \; + \; \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 \; + \; \eta^2 \operatorname{tr}(H) \right\}.$$

Moreover, for every $R \ge 0$,

$$2R^2 \operatorname{tr}(H_T) \; + \; \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{H_t^{-1}}^2 \le \left(\frac{R^2}{\eta^2} + \frac{1}{2} \right) \inf_{H \in \mathcal{H}} \left\{ \varepsilon \operatorname{tr}(H^{-1}) \; + \; \sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 \; + \; \eta^2 \operatorname{tr}(H) \right\}.$$

In particular, if $\eta \ge R$, then the coefficient on the right-hand side is at most $3/2$.

Proof of Lemma 10.4. Set

$$\ell_0(H) := \varepsilon \operatorname{tr}(H^{-1}) + \eta^2 \operatorname{tr}(H), \qquad \ell_t(H) := \|g_t\|_{H^{-1}}^2 \quad (t \ge 1).$$

Then H_t minimizes $L_t(H) := \sum_{s=0}^{t} \ell_s(H)$. We first prove the be-the-leader inequality

$$\sum_{t=0}^{T} \ell_t(H_t) \le \sum_{t=0}^{T} \ell_t(H) \qquad \forall H \in \mathcal{H}.$$

For $T = 0$, this is exactly the optimality of H_0 for L_0. Suppose the claim is known up to time $T - 1$. Apply the induction hypothesis with the comparator H_T:

$$\sum_{t=0}^{T-1} \ell_t(H_t) \le \sum_{t=0}^{T-1} \ell_t(H_T).$$

Adding $\ell_T(H_T)$ gives

$$\sum_{t=0}^{T} \ell_t(H_t) \le \sum_{t=0}^{T} \ell_t(H_T).$$

Since H_T minimizes L_T, the right-hand side is at most $\sum_{t=0}^{T} \ell_t(H)$ for every $H \in \mathcal{H}$. This proves the displayed be-the-leader inequality. Finally, $\ell_0(H_0) \ge 0$, so dropping it from the left-hand side and taking the infimum over H proves the first display in the lemma.

It remains to relate the terminal trace to the same raw objective. Since H_T minimizes the terminal objective and $\mathcal{H}$ is closed under positive scaling, scalar optimality at $c = 1$ for

$$c \mapsto \frac{1}{c}\left(\varepsilon \operatorname{tr}(H_T^{-1}) \; + \; \sum_{t=1}^{T} \|g_t\|^2_{H_T^{-1}}\right) \; + \; c\eta^2 \operatorname{tr}(H_T)$$

gives

$$\varepsilon \operatorname{tr}(H_T^{-1}) \; + \; \sum_{t=1}^{T} \|g_t\|^2_{H_T^{-1}} = \eta^2 \operatorname{tr}(H_T).$$

Therefore the terminal optimum value is $2\eta^2 \operatorname{tr}(H_T)$, and this terminal optimum value is the infimum in the lemma statement. Hence

$$2R^2 \operatorname{tr}(H_T) \le \frac{R^2}{\eta^2} \inf_{H \in \mathcal{H}} \left\{ \varepsilon \operatorname{tr}(H^{-1}) \; + \; \sum_{t=1}^{T} \|g_t\|^2_{H^{-1}} \; + \; \eta^2 \operatorname{tr}(H) \right\}.$$

Combining this trace bound with half of the first display gives the coefficient $R^2/\eta^2 + 1/2$. If $\eta \ge R$, this coefficient is at most $3/2$. □

Lemma 10.4 controls the inverse-metric gradient energy by the same comparator that motivated the algorithm. The remaining obstruction is metric growth: to use Theorem 10.3, the selected metrics must be increasing and their increments must lie in a cone on which the growth term is controlled. A general admissible family does not guarantee this, so we impose structure.

10.4 Well-structured Preconditioner Families

The role of the next definition is practical rather than decorative. We need a checkable condition on the admissible family $\mathcal{H}$ that guarantees three things simultaneously: the selector exists uniquely, the map $M \mapsto P_{\mathcal{H},\eta}(M)$ is order-preserving, and the increments of the selected metrics remain inside a fixed positive-semidefinite cone. Operator subalgebras give one such condition.

Definition 10.4 (Operator subalgebra)**.** A set $\mathcal{K} \subseteq \mathcal{L}(E)$ is called an operator subalgebra if

$$\forall \alpha \in \mathbb{R},\ \forall A, B \in \mathcal{K}, \qquad \alpha A \in \mathcal{K}, \quad A + B \in \mathcal{K}, \quad AB \in \mathcal{K}.$$

We always assume $I_E \in \mathcal{K}$.

Definition 10.5 (Well-structured preconditioner family and its closed cone)**.** A set $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ is called a *well-structured preconditioner family* if there exists an operator subalgebra $\mathcal{K} \subseteq \mathcal{L}(E)$ with $I_E \in \mathcal{K}$ such that

$$\mathcal{H} = \mathcal{K} \cap \mathcal{S}_{++}(E).$$

For such a family, write

$$\bar{\mathcal{H}} := \mathcal{K} \cap \mathcal{S}_{+}(E).$$

Definition 10.6 (Selected minimizer $P_{\mathcal{H},\eta}(M)$ and normalized witness)**.** Let $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ be a well-structured preconditioner family, let $\eta > 0$, and let $M \in \mathcal{S}_{++}(E)$. Whenever the

minimization problem

$$\arg\min_{H\in\mathcal{H}}\Big\{\operatorname{tr}(MH^{-1})+\eta^2\operatorname{tr}(H)\Big\}$$

is a singleton, we denote its unique element by $P_{\mathcal{H},\eta}(M)$. Its normalized version is

$$\widehat{P}_{\mathcal{H}}(M) := \frac{P_{\mathcal{H},\eta}(M)}{\operatorname{tr}(P_{\mathcal{H},\eta}(M))}.$$

This normalized operator is independent of the scaling parameter η, as shown in Proposition 10.5.

Proposition 10.5 (Structural facts for well-structured preconditioners)**.** *Let $\mathcal{H}\subseteq\mathcal{S}_{++}(E)$ be a well-structured preconditioner family, let $\bar{\mathcal{H}}$ be its closed cone, let $\eta>0$, and let $M\in\mathcal{S}_{++}(E)$. Then the selector problem defining $P_{\mathcal{H},\eta}(M)$ has a unique solution in $\mathcal{H}$. Moreover, the following properties hold.*

1. $$\operatorname{tr}(MP_{\mathcal{H},\eta}(M)^{-1})=\eta^2\operatorname{tr}(P_{\mathcal{H},\eta}(M)).$$

2. *The normalized operator $\widehat{P}_{\mathcal{H}}(M)$ is the unique element of*

$$\arg\min_{\substack{H\in\mathcal{H}\\ \operatorname{tr}(H)=1}}\operatorname{tr}(MH^{-1}).$$

3. *If $0\prec M\preceq M'$, then*

$$P_{\mathcal{H},\eta}(M)\preceq P_{\mathcal{H},\eta}(M') \qquad \text{and} \qquad P_{\mathcal{H},\eta}(M')-P_{\mathcal{H},\eta}(M)\in\bar{\mathcal{H}}.$$

Proof of Proposition 10.5. The existence, uniqueness, and monotonicity are imported from the structural preconditioner theorem, Proposition 3.2 of Xie–Wang–Reddi–Kumar–Li (2025). The proof of item 3 is the nontrivial subalgebra argument; it is not reproved here. For the reader's convenience, we record the two scalar computations that are used later in this lecture.
First, let $H=P_{\mathcal{H},\eta}(M)$. For every scalar $c>0$, the operator cH still belongs to $\mathcal{H}$, and optimality at $c=1$ for the one-variable function

$$c\mapsto\operatorname{tr}(M(cH)^{-1})+\eta^2\operatorname{tr}(cH)=\frac{1}{c}\operatorname{tr}(MH^{-1})+c\eta^2\operatorname{tr}(H)$$

forces

$$-\operatorname{tr}(MH^{-1})+\eta^2\operatorname{tr}(H)=0.$$

This proves item 1.
Second, for any $H\in\mathcal{H}$ with $\operatorname{tr}(H)>0$, write $\widehat{H}:=H/\operatorname{tr}(H)$. Then

$$\operatorname{tr}(MH^{-1})+\eta^2\operatorname{tr}(H)=\frac{\operatorname{tr}(M\widehat{H}^{-1})}{\operatorname{tr}(H)}+\eta^2\operatorname{tr}(H).$$

Minimizing first over the scalar $\operatorname{tr}(H)>0$ shows that the optimal value depends on $\widehat{H}$ only through $\operatorname{tr}(M\widehat{H}^{-1})$, and the minimizing scalar is

$$\operatorname{tr}(H)=\frac{1}{\eta}\sqrt{\operatorname{tr}(M\widehat{H}^{-1})}.$$

Hence minimizing the original objective over $\mathcal{H}$ is equivalent to minimizing $\operatorname{tr}(M\widehat{H}^{-1})$ over all $\widehat{H} \in \mathcal{H}$ with $\operatorname{tr}(\widehat{H}) = 1$, which proves item 2.
Item 3 is exactly the order-monotonicity and cone-increment statement in the same cited proposition. All later arguments in these notes use only the three displayed consequences above. □

The importance of Proposition 10.5 is immediate. Item 1 turns the chosen metric into a scalar complexity. Item 2 identifies the best static witness inside the family. Item 3 is exactly what the increasing-metric theorem needs: it shows that the selector output is monotone in M_t, and that the increments $H_t - H_{t-1}$ stay inside the same closed cone $\bar{\mathcal{H}}$.

The selector theorem below is stated with $\varepsilon > 0$, because the AdaReg matrices $M_t = \varepsilon I_E + \sum_{s=1}^{t}(g_s \otimes_E g_s)$ must stay positive definite in order for the selector $P_{\mathcal{H},\eta}(M_t)$ to be well defined at every step.

Main adaptive regret theorem. We can now return to the general increasing-metric theorem and specialize it to the AdaReg selector. The analysis separates metric growth from inverse-metric gradient energy.

Theorem 10.6 (Main adaptive regret guarantee for a well-structured family)**.** *Let $X \subseteq E$ be nonempty, closed, and convex. Let $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ be a well-structured preconditioner family, let $\bar{\mathcal{H}}$ be its closed cone, and fix $\eta > 0$ and $\varepsilon > 0$. Define*

$$M_0 := \varepsilon I_E, \qquad H_0 := P_{\mathcal{H},\eta}(M_0).$$

For every $t \in \{1, \dots, T\}$, let $L_t : X \to \mathbb{R}$ be convex, let $x_t \in X$, let $g_t \in \partial L_t(x_t)$, define

$$M_t := \varepsilon I_E + \sum_{s=1}^{t}(g_s \otimes_E g_s), \qquad H_t := P_{\mathcal{H},\eta}(M_t),$$

and let

$$x_{t+1} \in \arg\min_{x \in X}\Big\{\langle g_t, x\rangle_E + \frac{1}{2}\|x - x_t\|_{H_t}^2\Big\}.$$

Then, for every $u \in X$,

$$\sum_{t=1}^{T}(L_t(x_t) - L_t(u)) \le \left(\frac{\|X\|_{\mathcal{H}}^2}{\eta^2} + \frac{1}{2}\right) \inf_{H \in \mathcal{H}}\left\{\varepsilon \operatorname{tr}(H^{-1}) \; + \; \sum_{t=1}^{T}\|g_t\|_{H^{-1}}^2 \; + \; \eta^2 \operatorname{tr}(H)\right\}.$$

Moreover, if $d := \dim E$, then

$$\sum_{t=1}^{T}(L_t(x_t) - L_t(u)) \le \left(\frac{2\|X\|_{\mathcal{H}}^2}{\eta} + \eta\right)\left(|||g_{1:T}|||_{\mathcal{H}} + d\sqrt{\varepsilon}\right).$$

Proof of Theorem 10.6. Fix $u \in X$. Convexity gives

$$\sum_{t=1}^{T}(L_t(x_t) - L_t(u)) \le \sum_{t=1}^{T}\langle g_t, x_t - u\rangle_E.$$

If $\|X\|_{\mathcal{H}} = +\infty$, the displayed bounds are vacuous, so assume $\|X\|_{\mathcal{H}} < +\infty$. Since $M_1 \preceq \cdots \preceq$

M_T, item 3 of Proposition 10.5 gives $H_1 \preceq \cdots \preceq H_T$, $H_1 \in \mathcal{H} \subseteq \bar{\mathcal{H}}$, and $H_t - H_{t-1} \in \bar{\mathcal{H}}$ for $t \geq 2$. Applying the strengthened conclusion of Theorem 10.3, we get

$$\sum_{t=1}^{T}(L_t(x_t) - L_t(u)) \leq 2\,\|X\|_{\mathcal{H}}^2 \operatorname{tr}(H_T) \;+\; \frac{1}{2}\sum_{t=1}^{T}\|g_t\|_{H_t^{-1}}^2 .$$

The selected metrics satisfy the hypotheses of Lemma 10.4, because

$$\operatorname{tr}(M_t H^{-1}) = \varepsilon \operatorname{tr}(H^{-1}) + \sum_{s=1}^{t}\|g_s\|_{H^{-1}}^2 .$$

Applying that lemma with $R = \|X\|_{\mathcal{H}}$ proves the first displayed bound.
It remains to compare the raw infimum with the unregularized normalized complexity. Fix $K \in \mathcal{H}$ with $\operatorname{tr}(K) = 1$, and let

$$K_\alpha := \alpha K + (1-\alpha)\frac{I_E}{d}, \qquad \alpha \in (0,1).$$

Then $K_\alpha \in \mathcal{H}$, $\operatorname{tr}(K_\alpha) = 1$, and

$$K_\alpha^{-1} \preceq \alpha^{-1} K^{-1}, \qquad \operatorname{tr}(K_\alpha^{-1}) \leq \frac{d^2}{1-\alpha}.$$

Writing $A_K := \sum_{t=1}^{T}\|g_t\|_{K^{-1}}^2$, and using the candidate $H = rK_\alpha$, the raw infimum is at most

$$\inf_{r>0}\left\{\frac{A_K/\alpha + d^2\varepsilon/(1-\alpha)}{r} \;+\; \eta^2 r\right\} = 2\eta\sqrt{\frac{A_K}{\alpha} + \frac{d^2\varepsilon}{1-\alpha}}.$$

For fixed K, minimizing over $\alpha \in (0,1)$ gives

$$\inf_{\alpha\in(0,1)}\sqrt{\frac{A_K}{\alpha} + \frac{d^2\varepsilon}{1-\alpha}} = \sqrt{A_K} + d\sqrt{\varepsilon}.$$

Therefore

$$\inf_{H\in\mathcal{H}}\left\{\varepsilon \operatorname{tr}(H^{-1}) \;+\; \sum_{t=1}^{T}\|g_t\|_{H^{-1}}^2 \;+\; \eta^2 \operatorname{tr}(H)\right\} \leq 2\eta\left(\sqrt{A_K} + d\sqrt{\varepsilon}\right).$$

Taking the infimum over normalized K gives

$$\inf_{H\in\mathcal{H}}\left\{\varepsilon \operatorname{tr}(H^{-1}) \;+\; \sum_{t=1}^{T}\|g_t\|_{H^{-1}}^2 \;+\; \eta^2 \operatorname{tr}(H)\right\} \leq 2\eta\left(|||g_{1:T}|||_{\mathcal{H}} + d\sqrt{\varepsilon}\right).$$

Combining this with the first displayed bound proves the second displayed bound. □

Read online, the left-hand side of Theorem 10.6 is exactly the regret against comparator u. The second displayed bound is balanced by taking $\eta = \sqrt{2}\,\|X\|_{\mathcal{H}}$, which gives the coefficient $2\sqrt{2}\,\|X\|_{\mathcal{H}}$ in front of $|||g_{1:T}|||_{\mathcal{H}} + d\sqrt{\varepsilon}$. If the same convex objective f is used at every round, so that $L_t \equiv f$, then convexity gives

$$f\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right) - f(u) \leq \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f(u)),$$

so the same theorem immediately yields an offline nonsmooth bound for the averaged iterate.

10.5 Smooth Convex Objectives

The smooth case in the structured-preconditioner analysis is not a separate fixed-metric descent argument. It uses the regret theorem above, and then bounds the adaptive gradient complexity by a family-level smoothness constant.

Definition 10.7 (Adaptive smoothness)**.** Let $f : E \to \mathbb{R}$ be convex and twice continuously differentiable, and let $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ be a well-structured preconditioner family. Define

$$\mathsf{S}_{\mathcal{H}}(f) := \inf\Big\{\operatorname{tr}(A) : A \in \mathcal{H} \text{ and } \nabla^2 f(x) \preceq A \quad \forall x \in E\Big\}.$$

If no such A exists, set $\mathsf{S}_{\mathcal{H}}(f) = +\infty$.

This is the convex specialization of the adaptive smoothness definition of Xie–Wang–Reddi–Kumar–Li (2025). For nonconvex losses the paper uses the two-sided condition $-A \preceq \nabla^2 f(x) \preceq A$; for the convex optimization setting of these notes, the upper Hessian bound is the relevant part.

Theorem 10.7 (Smooth convergence rate of AdaReg)**.** *Let $X \subseteq E$ be nonempty, closed, convex, and bounded in $\|\cdot\|_{\mathcal{H}}$, with $\|X\|_{\mathcal{H}} > 0$, and let $d := \dim E$. Let $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ be a well-structured preconditioner family. Let $f : E \to \mathbb{R}$ be convex and twice continuously differentiable, suppose $x^\star \in X$ is a global minimizer of f, and suppose $\mathsf{S}_{\mathcal{H}}(f) < +\infty$. In the setup of Theorem 10.6, run Algorithm 1 with $L_t \equiv f$, $g_t = \nabla f(x_t)$, $\varepsilon > 0$, and*

$$\eta = \sqrt{2}\,\|X\|_{\mathcal{H}}\,.$$

If $\bar{x}_T := T^{-1}\sum_{t=1}^T x_t$, then

$$f(\bar{x}_T) - f(x^\star) \le \frac{16\,\|X\|_{\mathcal{H}}^2\,\mathsf{S}_{\mathcal{H}}(f)}{T} \;+\; \frac{4\sqrt{2}\,d\sqrt{\varepsilon}\,\|X\|_{\mathcal{H}}}{T}.$$

In the idealized $\varepsilon = 0$ display, this is the clean

$$f(\bar{x}_T) - f(x^\star) \le \frac{16\,\|X\|_{\mathcal{H}}^2\,\mathsf{S}_{\mathcal{H}}(f)}{T}.$$

Proof of Theorem 10.7. Set

$$R_T := \sum_{t=1}^T (f(x_t) - f(x^\star)), \qquad g_t := \nabla f(x_t), \qquad G_T := |||g_{1:T}|||_{\mathcal{H}}.$$

Fix any $A \in \mathcal{H}$ with $\nabla^2 f(x) \preceq A$ for all x, and set $H = A/\operatorname{tr}(A)$. The A-smooth upper model gives

$$f(x_t - A^{-1}g_t) \le f(x_t) - \frac{1}{2}\,\|g_t\|_{A^{-1}}^2\,.$$

Because $x^\star$ is a global minimizer, this implies

$$\|g_t\|_{A^{-1}}^2 \le 2\big(f(x_t) - f(x^\star)\big).$$

Since $H^{-1} = \operatorname{tr}(A)A^{-1}$, summing over t gives

$$\sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2 \le 2\operatorname{tr}(A) R_T.$$

Taking the infimum over normalized H, and then over admissible A, yields

$$G_T^2 \le 2\,\mathsf{S}_{\mathcal{H}}(f) R_T.$$

Using the second bound in Theorem 10.6 with $u = x^\star$ and $\eta = \sqrt{2}\,\|X\|_{\mathcal{H}}$ gives

$$R_T \le 2\sqrt{2}\,\|X\|_{\mathcal{H}}\,(G_T + d\sqrt{\varepsilon}) \le 4\,\|X\|_{\mathcal{H}}\sqrt{\mathsf{S}_{\mathcal{H}}(f) R_T} \;+\; 2\sqrt{2}\,d\sqrt{\varepsilon}\,\|X\|_{\mathcal{H}}\,.$$

Let $y := \sqrt{R_T}$. Then $y^2 \le ay + b$, where

$$a = 4\,\|X\|_{\mathcal{H}}\sqrt{\mathsf{S}_{\mathcal{H}}(f)}, \qquad b = 2\sqrt{2}\,d\sqrt{\varepsilon}\,\|X\|_{\mathcal{H}}\,.$$

Since $y^2 \le ay + b$ implies $y^2 \le a^2 + 2b$,

$$R_T \le 16\,\|X\|_{\mathcal{H}}^2\;\mathsf{S}_{\mathcal{H}}(f) \;+\; 4\sqrt{2}\,d\sqrt{\varepsilon}\,\|X\|_{\mathcal{H}}\,.$$

Finally, Jensen's inequality gives

$$f(\bar{x}_T) - f(x^\star) \le \frac{1}{T} R_T,$$

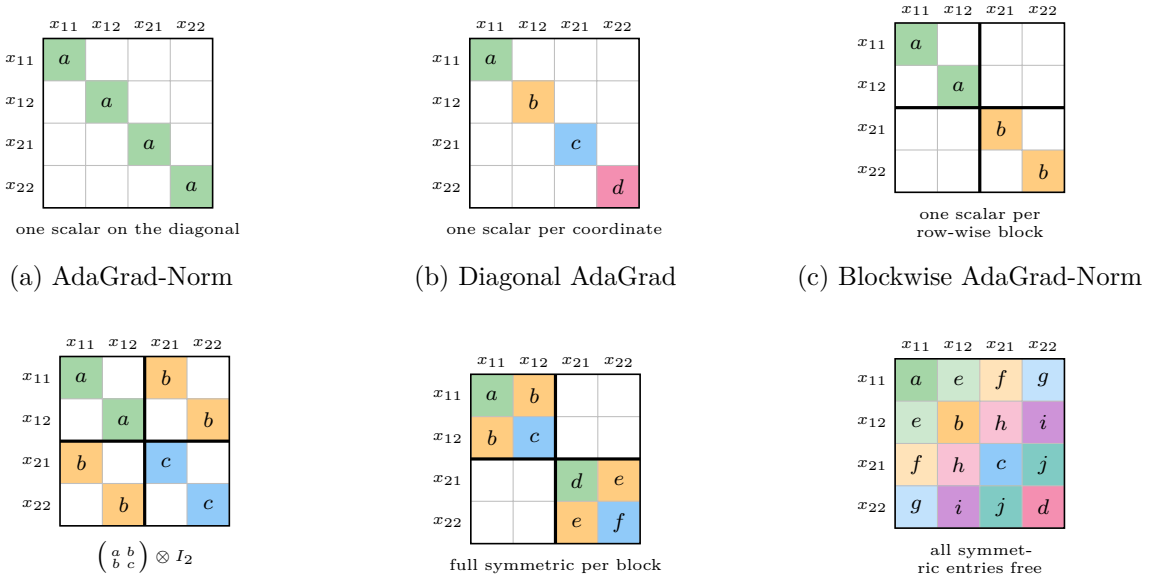
which proves the theorem. □

10.6 Examples: Canonical Well-structured Families

The theorem above treats the selector abstractly. The point of the examples below is to identify the most common adaptive optimizers as particular structured families. From this point on, choose an orthonormal basis of E and identify $E \cong \mathbb{R}^d$, where $d := \dim E$. Under this identification, $\mathcal{L}(E)$ becomes $\mathbb{R}^{d\times d}$, and $\mathcal{S}_{++}(E)$ becomes $\mathbb{S}^d_{++}$. The families below are basis-dependent by design. When $d = d_L d_R$ and a vector $x \in \mathbb{R}^d$ is viewed as a matrix $X \in \mathbb{R}^{d_L \times d_R}$, we use the row-wise vectorization[1]

$$\operatorname{rvec}(X) := \begin{pmatrix} X_{11} & \cdots & X_{1d_R} & X_{21} & \cdots & X_{2d_R} & \cdots & X_{d_L 1} & \cdots & X_{d_L d_R} \end{pmatrix}^\top \in \mathbb{R}^d.$$

[1]Here $g \otimes_E g$ in the basis-free part denotes the rank-one operator $x \mapsto \langle g, x\rangle_E\, g$, while $A \otimes B$ in this coordinate-dependent part denotes the Kronecker product; for row-wise vectorization, $(A \otimes I_{d_R})\operatorname{rvec}(X) = \operatorname{rvec}(AX)$.



(a) AdaGrad-Norm (b) Diagonal AdaGrad (c) Blockwise AdaGrad-Norm

(d) Left-sided Shampoo (e) Block-diag. Full AdaGrad (f) Full-matrix AdaGrad

Figure 1: Parameter-sharing structure of six adaptive preconditioner families. White entries are forced zeros; the picture shows admissible structure, not a typical realized matrix. Block-diagonal full AdaGrad (e) interpolates between blockwise AdaGrad-Norm (c) and full-matrix AdaGrad (f); it has no standard name in the literature.

Panel c is called blockwise AdaGrad-Norm in these notes.[2]

Update-rule convention. In the six examples below, the displayed update rules are the unconstrained forms

$$x_{t+1} = x_t - H_t^{-1} g_t.$$

For a constrained feasible set X, replace each displayed rule by the quadratic proxy step in Algorithm 1. Thus projection is omitted only to show the optimizer's familiar coordinate formula.

Example 10.1 (AdaGrad-Norm)**.** The scalar family is

$$\mathcal{H}_{\text{scalar}} := \{cI_d : c > 0\}.$$

The selected metric is

$$H_t = \frac{1}{\eta}\sqrt{\varepsilon + \frac{1}{d}\sum_{s=1}^{t} \|g_s\|_2^2}\, I_d.$$

Hence

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{\varepsilon + \frac{1}{d}\sum_{s=1}^{t} \|g_s\|_2^2}}\, g_t.$$

[2] In machine-learning usage, "layerwise AdaGrad" often means one scalar per layer; this is blockwise AdaGrad-Norm, not diagonal AdaGrad restricted to each layer.

Example 10.2 (Diagonal AdaGrad)**.** The diagonal family is

$$\mathcal{H}_{\mathrm{diag}} := \left\{ \mathrm{Diag}(h) : h \in \mathbb{R}^d_{++} \right\}.$$

For $g_s = (g_{s,1}, \dots, g_{s,d})$, the selected metric is

$$H_t = \frac{1}{\eta} \mathrm{Diag}\left(\sqrt{\varepsilon + \sum_{s=1}^t g_{s,1}^2}, \dots, \sqrt{\varepsilon + \sum_{s=1}^t g_{s,d}^2} \right).$$

Hence

$$x_{t+1,i} = x_{t,i} - \frac{\eta\, g_{t,i}}{\sqrt{\varepsilon + \sum_{s=1}^t g_{s,i}^2}} \qquad (i = 1, \dots, d).$$

Example 10.3 (Blockwise AdaGrad-Norm)**.** Suppose $d = d_1 + \cdots + d_m$, write $x = (x_1, \dots, x_m)$ and $g_t = (g_{t,1}, \dots, g_{t,m})$ with $x_\ell, g_{t,\ell} \in \mathbb{R}^{d_\ell}$, and let

$$\mathcal{H}_{\mathrm{block}} := \{ h_1 I_{d_1} \oplus \cdots \oplus h_m I_{d_m} : h_1, \dots, h_m > 0 \}.$$

The selected metric is

$$H_t = \frac{1}{\eta} \bigoplus_{\ell=1}^m \sqrt{\varepsilon + \frac{1}{d_\ell} \sum_{s=1}^t \|g_{s,\ell}\|_2^2}\, I_{d_\ell}.$$

Hence

$$x_{t+1,\ell} = x_{t,\ell} - \frac{\eta}{\sqrt{\varepsilon + \frac{1}{d_\ell} \sum_{s=1}^t \|g_{s,\ell}\|_2^2}}\, g_{t,\ell} \qquad (\ell = 1, \dots, m).$$

Example 10.4 (Left-sided Shampoo)**.** Write $d = d_L d_R$, reshape $x_t = \mathrm{rvec}(X_t)$, and reshape $g_t = \mathrm{rvec}(G_t)$, where $X_t, G_t \in \mathbb{R}^{d_L \times d_R}$. The left-sided Shampoo family is

$$\mathcal{H}_{\mathrm{left}} := \left\{ H_L \otimes I_{d_R} : H_L \in \mathbb{S}^{d_L}_{++} \right\}.$$

The selected metric has the form

$$H_t = \frac{1}{\eta} \left(\varepsilon I_{d_L} + \frac{1}{d_R} \sum_{s=1}^t G_s G_s^\top \right)^{1/2} \otimes I_{d_R}.$$

Using $(A \otimes I_{d_R})\, \mathrm{rvec}(X) = \mathrm{rvec}(AX)$, the update rule is

$$X_{t+1} = X_t - \eta \left(\varepsilon I_{d_L} + \frac{1}{d_R} \sum_{s=1}^t G_s G_s^\top \right)^{-1/2} G_t.$$

This update is called one-sided Shampoo in Xie–Wang–Reddi–Kumar–Li (2025) and ASGO in An–Liu–Pan–Ma–Goldfarb–Zhang (2025). It should be distinguished from the original two-sided Shampoo method of [GKS18], which uses both left and right matrix factors. For smooth objectives on matrices, this same family measures left curvature: the quantity $\mathsf{S}_{\mathcal{H}}(f)$ is the smallest number of the form $d_R \,\mathrm{tr}(A_L)$ such that

$$\nabla^2 f(X)[\Delta, \Delta] \le \mathrm{tr}(A_L \Delta \Delta^\top) \qquad \forall X, \Delta \in \mathbb{R}^{d_L \times d_R}.$$

This is the left-smoothness constant used in the one-sided Shampoo specialization of the paper.

Example 10.5 (Block-diagonal full AdaGrad)**.** Suppose $d = d_1 + \cdots + d_m$, and write $x = (x_1, \ldots, x_m)$ and $g_t = (g_{t,1}, \ldots, g_{t,m})$ with $x_\ell, g_{t,\ell} \in \mathbb{R}^{d_\ell}$. The block-diagonal full family is

$$\mathcal{H}_{\mathrm{bd}} := \left\{ H_1 \oplus \cdots \oplus H_m : H_\ell \in \mathbb{S}^{d_\ell}_{++} \right\}.$$

The selected metric is

$$H_t = \frac{1}{\eta} \bigoplus_{\ell=1}^{m} \left(\varepsilon I_{d_\ell} + \sum_{s=1}^{t} g_{s,\ell} g_{s,\ell}^\top \right)^{1/2}.$$

Hence

$$x_{t+1,\ell} = x_{t,\ell} - \eta \left(\varepsilon I_{d_\ell} + \sum_{s=1}^{t} g_{s,\ell} g_{s,\ell}^\top \right)^{-1/2} g_{t,\ell} \qquad (\ell = 1, \ldots, m).$$

This family interpolates between blockwise AdaGrad-Norm and full-matrix AdaGrad; it is included to explain the structure in Figure 1, not because it has a standard optimizer name.

Example 10.6 (Full-matrix AdaGrad)**.** The full family is

$$\mathcal{H}_{\mathrm{full}} := \mathbb{S}^d_{++}.$$

The selected metric is

$$H_t = \frac{1}{\eta} \left(\varepsilon I_d + \sum_{s=1}^{t} g_s g_s^\top \right)^{1/2}.$$

Hence

$$x_{t+1} = x_t - \eta \left(\varepsilon I_d + \sum_{s=1}^{t} g_s g_s^\top \right)^{-1/2} g_t.$$

Proof of Example 10.1–Example 10.6. Each displayed family is the intersection of $\mathbb{S}^d_{++}$ with a matrix subalgebra that contains the identity: scalar matrices, diagonal matrices, block-scalar direct sums, left Kronecker products, block-diagonal matrix direct sums, and the full matrix algebra are each closed under scalar multiplication, addition, and multiplication. Hence every family is well structured.

The scalar, diagonal, and block-scalar formulas follow by minimizing one-dimensional terms of the form $a/h + \eta^2 bh$. For the full family, the first-order condition for $\mathrm{tr}(MH^{-1}) + \eta^2 \mathrm{tr}(H)$ is $H^{-1}MH^{-1} = \eta^2 I_d$, hence $H = \eta^{-1} M^{1/2}$. The block-diagonal full family is the same calculation on each block. For left-sided Shampoo, the identity

$$\mathrm{rvec}(G)^\top (H_L^{-1} \otimes I_{d_R}) \, \mathrm{rvec}(G) = \mathrm{tr}(GG^\top H_L^{-1})$$

reduces the selector to the full-matrix calculation on the left factor, with cumulative matrix $\varepsilon I_{d_L} + d_R^{-1} \sum_{s=1}^{t} G_s G_s^\top$. Substituting $M_t = \varepsilon I_d + \sum_{s=1}^{t} g_s g_s^\top$ gives the displayed metrics, and each update is the unconstrained proxy step $x_{t+1} = x_t - H_t^{-1} g_t$. □

Dependency and Proof Sketch

1. Corollary 10.1 is the fixed-geometry consequence of Lecture 9. It motivates the hindsight problem of choosing the best single $H \in \mathcal{H}$ after seeing the gradients.

2. Lemma 10.2 is the fixed-metric quadratic mirror-descent inequality used inside the general adaptive-metric analysis.

3. Theorem 10.3 is the master theorem for arbitrary increasing metrics. It identifies the two quantities any adaptive rule must control: metric growth and inverse-metric gradient energy.

4. Algorithm 1 is the concrete online rule for choosing H_t: it replaces the unknown terminal second-moment operator by the observed prefix M_t, chooses a metric H_t, and then takes the quadratic proxy step. By Theorem 10.3, its analysis hinges on monotonicity of the selected metrics.

5. Proposition 10.5 is the only imported structural result. Its role is not merely to define the selector, but to guarantee monotonicity and cone membership of the increments $H_t - H_{t-1}$.

6. Theorem 10.6 combines three ingredients: the increasing-metric bound Theorem 10.3, the be-the-leader control of the prefix selector in Lemma 10.4, and the structural monotonicity / normalization facts in Proposition 10.5. This is the main theorem of the lecture: AdaReg is analyzed by separating metric growth from inverse-metric gradient energy, then showing that the well-structured selector controls both.

7. Definition 10.7 and Theorem 10.7 give the deterministic smooth case from the structured-preconditioner paper: family smoothness controls the adaptive gradient complexity, and the regret theorem then becomes a $1/T$ smooth convergence rate.

8. Example 10.1, Example 10.2, Example 10.3, Example 10.4, Example 10.5, and Example 10.6 place the standard AdaGrad-style update rules inside one preconditioner framework rather than treating them as separate analyses.

Exercises

1. Starting from Theorem 10.6, specialize the adaptive regret bound to the following three families:

$$\mathcal{H}_{\text{diag}}, \qquad \mathcal{H}_{\text{left}} = \{H_L \otimes I_{d_R} : H_L \succ 0\}, \qquad \mathcal{H}_{\text{full}} = \mathbb{S}^d_{++}.$$

   For each family, write the unregularized normalized complexity $|||g_{1:T}|||_{\mathcal{H}}$ explicitly in terms of the observed gradients, and then write the resulting regret bound after choosing $\eta = \|X\|_{\mathcal{H}}$.

2. The original Shampoo algorithm [GKS18] uses two-sided matrix preconditioning for matrix gradients $G_t \in \mathbb{R}^{d_L \times d_R}$:

$$L_t := \left(\varepsilon I_{d_L} + \sum_{s=1}^{t} G_s G_s^\top\right)^{1/4}, \qquad R_t := \left(\varepsilon I_{d_R} + \sum_{s=1}^{t} G_s^\top G_s\right)^{1/4},$$

   and the unconstrained update

$$X_{t+1} = X_t - \eta L_t^{-1} G_t R_t^{-1}.$$

Under the row-wise vectorization convention, identify the induced vector-space metric H_t. Is this two-sided Shampoo rule an instance of the AdaReg selector $P_{\mathcal{H},\eta}(M_t)$ for a well-structured family $\mathcal{H}$ from this lecture? Does the sequence H_t satisfy Loewner monotonicity? Finally, explain why Loewner monotonicity alone is not the same as the increment condition $H_t - H_{t-1} \in \bar{\mathcal{H}}$ used in Theorem 10.6.

Historical and Bibliographic Notes

The adaptive-gradient line includes the online-conditioning view of [SM10] and the AdaGrad framework of [DHS11]; see also the adaptive-bound perspective of [MS10]. Adam [KB15] adds exponential moving averages of first and second moments together with bias correction; it is central in practice, but its momentum and exponential forgetting make it a different object from the monotone-prefix selector analyzed in this lecture. The AdaReg template itself comes from [GKS17]. Shampoo [GKS18] introduced matrix and tensor preconditioning through Kronecker-structured second-moment factors. The one-sided update in Example 10.4 appears as ASGO in An–Liu–Pan–Ma–Goldfarb–Zhang (2025) and as one-sided Shampoo in the well-structured-preconditioner analysis of Xie–Wang–Reddi–Kumar–Li (2025).

References

[ALP+25] Kang An, Yuxing Liu, Rui Pan, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[GKS17] Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in online and stochastic optimization. *arXiv preprint arXiv:1706.06569*, 2017.

[GKS18] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 2018.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[MS10] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, 2010.

[SM10] Matthew Streeter and H. Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

[XWR+25] Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. *arXiv preprint arXiv:2503.10537*, 2025.