Lecture 10: Adaptive Optimization and Well-structured Preconditioners

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

From Fixed Geometry to an Adaptive Target

Lecture 9: one mirror map Φ fixed for the whole run. **Today:** choose the geometry H_t online, from past gradients.

Quadratic preconditioner geometry. For $H \in \mathcal{S}_{++}(E)$, $\Phi_H(x) := \frac{1}{2}\langle x, Hx\rangle_E$, so $D_{\Phi_H}(x, y) = \frac{1}{2}\|x - y\|_H^2$. The mirror step is

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

$\Rightarrow$ choosing a quadratic mirror map = choosing a positive-definite preconditioner H. The question is how to choose it.

Fixed-Metric Regret

Corollary 10.1. For $H \in \mathcal{S}_{++}(E)$ and the quadratic proxy step with $\eta_t \equiv 1$,

$$\sum_{t=1}^{T} \langle g_t, x_t - u \rangle_E \;\le\; \underbrace{\tfrac{1}{2}\|u - x_1\|_H^2}_{\text{init. distance}} + \underbrace{\tfrac{1}{2}\sum_{t=1}^{T} \|g_t\|_{H^{-1}}^2}_{\text{inverse-metric gradient energy}} .$$

Writing $\|X\|_H := \sup_{x \in X} \|x\|_H$ and using $\|u - x_1\|_H \le \|u\|_H + \|x_1\|_H \le 2\|X\|_H$:

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

Tension in H: larger H shrinks gradient energy but inflates $\|X\|_H$; smaller H does the opposite.

Proof. Cor 9.4 on Φ_H: 1-SC w.r.t. $\|\cdot\|_H$, dual $\|\cdot\|_{H^{-1}}$.

Hindsight: Best Fixed H and the Family $\mathcal{H}$

Best hindsight H (knowing $g_{1:T}$ in advance) minimizes $2\|X\|_H^2 + \frac{1}{2}\sum_t \|g_t\|_{H^{-1}}^2$ over $\mathcal{H} \subseteq \mathcal{S}_{++}(E)$ (closed under positive scaling).

Family norm $\|x\|_{\mathcal{H}} := \sup_{H \in \overline{\mathcal{H}},\, \mathrm{tr}(H) \le 1} \sqrt{\langle x, Hx\rangle_E}$ — largest H-seminorm of x over unit-trace shapes in the family.
Domain radius $\|X\|_{\mathcal{H}} := \sup_{x \in X} \|x\|_{\mathcal{H}}$.

Product form of hindsight optimum. Write $H = r\widehat{H}$ with $\mathrm{tr}\,\widehat{H} = 1$; minimize over r:

$$\inf_{H \in \mathcal{H}}(\cdot) = 2\|X\|_{\mathcal{H}} \cdot \inf_{\widehat{H} \in \mathcal{H},\, \mathrm{tr}\,\widehat{H}=1} \sqrt{\textstyle\sum_t \|g_t\|_{\widehat{H}^{-1}}^2}.$$

$=$ *twice domain radius* $\times$ *best normalized gradient energy*.

Adaptive Proxy Update and One-Step Inequality

Definition. Adaptive quadratic proxy update: for $H_t \in \mathcal{S}_{++}(E)$,

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

Lemma 10.2 (One-step inequality). For any $H \in \mathcal{S}_{++}(E)$, $x_t, u \in X$, $g_t \in E$, and the above update,

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

Proof. Quadratic mirror step + Thm 8.9 (one-step) + Lem 9.3 (Young in $\|\cdot\|_H, \|\cdot\|_{H^{-1}}$).

Increasing-Metric Comparison Bound

Theorem 10.3. If $H_1 \preceq H_2 \preceq \cdots \preceq H_T$ in $\mathcal{S}_{++}(E)$:

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

If additionally $H_1 \in \overline{\mathcal{H}}$ and $H_t - H_{t-1} \in \overline{\mathcal{H}}$ for all $t \geq 2$ (with $\mathcal{H}$ a family closed under positive scaling), then:

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

Reading. Two budgets: (a) *metric growth* $\operatorname{tr}(H_T)$ (domain cost); (b) *inverse-metric energy* $\sum \|g_t\|_{H_t^{-1}}^2$ (gradient cost). Adaptive rule must control both.

Thm 10.3: Proof

Step 1. Apply Lem 10.2 at each t with same u; sum:

$$\sum_t \langle g_t, x_t - u\rangle_E \le \tfrac{1}{2}\sum_t [\|u - x_t\|_{H_t}^2 - \|u - x_{t+1}\|_{H_t}^2] + \tfrac{1}{2}\sum_t \|g_t\|_{H_t^{-1}}^2.$$

Step 2. Telescope with changing metric. Re-index the first sum: $\|u - x_t\|_{H_t}^2 - \|u - x_{t+1}\|_{H_t}^2$ across consecutive t gives $\|u - x_1\|_{H_1}^2 + \sum_{t\ge 2}(\|u - x_t\|_{H_t}^2 - \|u - x_t\|_{H_{t-1}}^2)$
$= \|u - x_1\|_{H_1}^2 + \sum_{t\ge 2}\|u - x_t\|_{H_t - H_{t-1}}^2$ (drops nonpositive $-\|u - x_{T+1}\|_{H_T}^2$). First inequality.

Step 3. $A \in \overline{\mathcal{H}} \Rightarrow \|z\|_A^2 = \mathrm{tr}(A)\|z\|_{A/\mathrm{tr}\,A}^2 \le \mathrm{tr}(A)\|z\|_{\mathcal{H}}^2$. With $u, x_t \in X$: $\|u - x_t\|_{\mathcal{H}} \le 2\|X\|_{\mathcal{H}}$: $\|u - x_1\|_{H_1}^2 \le 4\|X\|_{\mathcal{H}}^2\,\mathrm{tr}(H_1)$, $\|u - x_t\|_{H_t - H_{t-1}}^2 \le 4\|X\|_{\mathcal{H}}^2\,\mathrm{tr}(H_t - H_{t-1})$.

Step 4. Sum: $\mathrm{tr}(H_1) + \sum_{t\ge 2}\mathrm{tr}(H_t - H_{t-1}) = \mathrm{tr}(H_T)$. $\square$

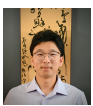
AdaReg: Choosing the Metric Online

Idea. Replace the unknown terminal gradient sum in the hindsight objective by its current prefix. *Be-the-leader* rule.

Cumulative gradient covariance: $(g \otimes_E g)(x) := \langle g, x\rangle_E\, g$ (coord.: $gg^\top$); $M_0 := \varepsilon I_E$, $M_t := M_{t-1} + g_t \otimes_E g_t$.

AdaReg meta-algorithm. Parameters $\eta, \varepsilon > 0$, family $\mathcal{H}$, $x_1 \in X$. For $t = 1, 2, \ldots, T$: (i) observe g_t; update M_t; (ii) select $H_t \in \operatorname{argmin}_{H\in\mathcal{H}}\{\operatorname{tr}(M_t H^{-1}) + \eta^2 \operatorname{tr}(H)\}$; (iii) step $x_{t+1} \in \operatorname{argmin}_{x\in X}\{\langle g_t, x\rangle_E + \frac{1}{2}\|x - x_t\|^2_{H_t}\}$.


Shuo Xie


Tianhao Wang


Sashank Reddi


Sanjiv Kumar


Zhiyuan Li

Be-the-Leader for the Selector

Why this objective? $\operatorname{tr}(M_t H^{-1})$ penalizes H small where past gradients were large; $\eta^2 \operatorname{tr}(H)$ penalizes H too large overall. Prefix version of the fixed-metric tradeoff.

Lemma 10.4 (Be-the-leader). Let H_t minimize the prefix objective through step t. Then

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

If $\mathcal{H}$ is closed under positive scaling, the RHS equals $2\eta \cdot \Gamma_{\mathcal{H}}(g_{1:T}; \varepsilon)$ where $\Gamma_{\mathcal{H}}(g_{1:T}; \varepsilon) := \inf_{\widehat{H} \in \mathcal{H}, \operatorname{tr} \widehat{H} = 1} \sqrt{\varepsilon \operatorname{tr} \widehat{H}^{-1} + \sum_t \|g_t\|^2_{\widehat{H}^{-1}}}$.

Proof. Set $\ell_0(H) := \varepsilon \operatorname{tr}(H^{-1}) + \eta^2 \operatorname{tr}(H)$, $\ell_t(H) := \|g_t\|^2_{H^{-1}}$. H_t minimizes $\sum_{s \le t} \ell_s$. BTL induction: $\sum_t \ell_t(H_t) \le \sum_t \ell_t(H)$ for all H. Drop $\ell_0(H_0) \ge 0$; take inf. Scaling: min over $r = \operatorname{tr} H$ of $\frac{A}{r} + \eta^2 r$ is $2\eta\sqrt{A}$. $\square$

Well-Structured Preconditioner Families

Why extra structure? Thm 10.3 needs increments $H_t - H_{t-1}$ to stay in a fixed cone. General families don't guarantee this. Operator subalgebras do.

Operator subalgebra $\mathcal{K} \subseteq \mathcal{L}(E)$: closed under $\alpha A, A + B, AB$ (any $\alpha \in \mathbb{R}$, $A, B \in \mathcal{K}$); contains I_E.

Well-structured preconditioner family. $\mathcal{H} = \mathcal{K} \cap \mathcal{S}_{++}(E)$ for some operator subalgebra $\mathcal{K} \ni I_E$.
Closed cone: $\overline{\mathcal{H}} = \mathcal{K} \cap \mathcal{S}_{+}(E)$.

Selected minimizer. $P_{\mathcal{H},\eta}(M) :=$ unique minimizer of $H \mapsto \operatorname{tr}(MH^{-1}) + \eta^2 \operatorname{tr}(H)$ over $\mathcal{H}$. Normalized: $\widehat{P}_{\mathcal{H}}(M) := P_{\mathcal{H},\eta}(M) / \operatorname{tr} P_{\mathcal{H},\eta}(M)$ (η-independent).

Properties of Well-Structured Preconditioners

Proposition 10.5 (Xie–Wang–Reddi–Kumar–Li, 2025). For well-structured $\mathcal{H}$, $\eta > 0$, $M \in \mathcal{S}_{++}(E)$: (i) *trace balance* $\operatorname{tr}(MP_{\mathcal{H},\eta}(M)^{-1}) = \eta^2 \operatorname{tr}(P_{\mathcal{H},\eta}(M))$; (ii) *normalized optimum* $\widehat{P}_{\mathcal{H}}(M) = \operatorname{argmin}\{\operatorname{tr}(MH^{-1}) : H \in \mathcal{H}, \operatorname{tr} H = 1\}$; (iii) *order-preserving with cone increments* $0 \prec M \preceq M' \Rightarrow P_{\mathcal{H},\eta}(M) \preceq P_{\mathcal{H},\eta}(M')$, $P_{\mathcal{H},\eta}(M') - P_{\mathcal{H},\eta}(M) \in \overline{\mathcal{H}}$.

Role: (i) $\to$ scalar complexity $\eta^2 \operatorname{tr}(H_T)^2 = \Gamma_{\mathcal{H}}(g_{1:T}; \varepsilon)^2$; (ii) $\to$ static witness in the family; (iii) $\to$ feeds Thm 10.3.

Shuo Xie

Tianhao Wang

Sashank Reddi

Sanjiv Kumar

Zhiyuan Li

Main Adaptive Regret Guarantee

Theorem 10.6 (Xie–Wang–Reddi–Kumar–Li, 2025). Well-structured $\mathcal{H}$; $\eta, \varepsilon > 0$; $H_t = P_{\mathcal{H},\eta}(M_t)$; convex L_t with $g_t \in \partial L_t(x_t)$; adaptive proxy step. Then for all $u \in X$:

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

where $d = \dim E$ and $|||g_{1:T}|||_{\mathcal{H}} := \inf_{\widehat{H} \in \mathcal{H}, \mathrm{tr}=1} \sqrt{\sum_t \|g_t\|_{\widehat{H}^{-1}}^2}$.

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

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

Matches the hindsight product form up to constants!

Thm 10.6: Proof

Step 1. Convexity: $L_t(x_t) - L_t(u) \le \langle g_t, x_t - u\rangle_E$.

Step 2. $M_1 \preceq \cdots \preceq M_T$, so by Prop 10.5(iii) $H_1 \preceq \cdots \preceq H_T$ with increments in $\overline{\mathcal{H}}$. Thm 10.3 (strengthened): $\sum\langle g_t, x_t - u\rangle_E \le 2\|X\|_{\mathcal{H}}^2 \operatorname{tr}(H_T) + \frac{1}{2}\sum\|g_t\|_{H_t^{-1}}^2$.

Step 3. Lem 10.4 (BTL): $\sum\|g_t\|_{H_t^{-1}}^2 \le 2\eta\Gamma_{\mathcal{H}}(g_{1:T};\varepsilon)$.

Step 4. Prop 10.5(i)+(ii): $\eta^2\operatorname{tr}(H_T)^2 = \Gamma_{\mathcal{H}}(g_{1:T};\varepsilon)^2$, so $\operatorname{tr}(H_T) = \Gamma_{\mathcal{H}}/\eta$. Hence $2\|X\|_{\mathcal{H}}^2\operatorname{tr}(H_T) + \eta\Gamma_{\mathcal{H}} = \big(\frac{2\|X\|_{\mathcal{H}}^2}{\eta} + \eta\big)\Gamma_{\mathcal{H}}$.

Step 5. $\Gamma_{\mathcal{H}} \le |||g_{1:T}|||_{\mathcal{H}} + d\sqrt{\varepsilon}$ (interpolate $\widehat{H}$ with I_E/d, optimize $\alpha \in (0,1)$). $\square$

$\mathcal{H}$-Smoothness

Definition 10.7 ($\mathcal{H}$-smoothness). For $f : E \to \mathbb{R}$ convex, C^2, and well-structured $\mathcal{H}$, set

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

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

Reading. Family-level generalization of scalar L-smoothness.

- $\mathcal{H} = \{cI\}$: $\mathsf{S} = d \cdot L$ where $L = \sup_x \lambda_{\max}(\nabla^2 f(x))$.
- $\mathcal{H} = \text{diag}$: $\mathsf{S} = \sum_i \sup_x |\partial_{ii}^2 f(x)|$ (coordinatewise).
- $\mathcal{H} = \mathbb{S}_{++}$: $\mathsf{S} = \inf_{A \succeq \nabla^2 f} \operatorname{tr}(A)$ (full).

Matching the geometry to the Hessian reduces $\mathsf{S}_{\mathcal{H}}$ dramatically.

Smooth Convergence of AdaReg

Theorem 10.7. f convex, C^2; $\mathsf{S}_{\mathcal{H}}(f) < \infty$; $x^\star \in X$ global min; run AdaReg with $L_t \equiv f$, $g_t = \nabla f(x_t)$, $\eta = \sqrt{2}\|X\|_{\mathcal{H}}$. Then for $\bar{x}_T = \frac{1}{T}\sum x_t$:

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

At $\varepsilon \to 0$: **clean** $O(\|X\|_{\mathcal{H}}^2\,\mathsf{S}_{\mathcal{H}}(f)/T)$ **smooth rate**.

Proof sketch. (1) *Self-bounding.* A-smooth upper model: $f(x_t - A^{-1}g_t) \le f(x_t) - \frac{1}{2}\|g_t\|_{A^{-1}}^2$; since $f(x_t - A^{-1}g_t) \ge f(x^\star)$: $\|g_t\|_{A^{-1}}^2 \le 2(f(x_t) - f(x^\star))$. (2) Normalize $H = A/\operatorname{tr} A$; sum and inf over A: $|||g_{1:T}|||_{\mathcal{H}}^2 \le 2\mathsf{S}_{\mathcal{H}}(f)\,R_T$, $R_T := \sum(f(x_t) - f(x^\star))$. (3) Thm 10.6 + step (2): $R_T \le 2\sqrt{2}\|X\|_{\mathcal{H}}\big(\sqrt{2\mathsf{S}R_T} + d\sqrt{\varepsilon}\big)$. (4) $y = \sqrt{R_T}$ satisfies $y^2 \le ay + b$ with $a = 4\|X\|_{\mathcal{H}}\sqrt{\mathsf{S}}$, $b = 2\sqrt{2}d\sqrt{\varepsilon}\|X\|_{\mathcal{H}} \Rightarrow y \le a + \sqrt{b} \Rightarrow R_T \le 2a^2 + 2b$. Jensen: $f(\bar{x}_T) - f(x^\star) \le R_T/T$. $\square$

Canonical Well-Structured Families

Family	Structure	Parameters
AdaGrad-Norm	cI_d, $c > 0$	1 scalar
Diagonal AdaGrad	$\mathrm{diag}(h)$, $h \in \mathbb{R}^d_{++}$	d scalars
Blockwise AdaGrad-Norm	$\bigoplus h_\ell I_{d_\ell}$	m block scalars
Left-sided Shampoo	$H_L \otimes I_{d_R}$	$d_L(d_L+1)/2$
Block-diag. full AdaGrad	$\bigoplus H_\ell$	$\sum d_\ell(d_\ell+1)/2$
Full-matrix AdaGrad	$\mathcal{S}^d_{++}$	$d(d+1)/2$

Each is the intersection of $\mathcal{S}^d_{++}$ with a matrix subalgebra containing I: scalar matrices, diagonals, block-scalar sums, left Kronecker, block diagonals, full matrices. All are *well-structured*.

Update rules (unconstrained form $x_{t+1} = x_t - H_t^{-1} g_t$): the selected H_t takes closed form in each case (next slides).

AdaGrad-Norm

Example 10.1 (AdaGrad-Norm; McMahan–Streeter, COLT 2010; named by Ward–Wu–Bottou, ICML 2019). $\mathcal{H} = \{cI_d : c > 0\}$. Selected metric

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

One global scalar stepsize that decays with cumulative gradient energy.

H. B. McMahan

Matthew Streeter

Diagonal AdaGrad

Example 10.2 (Diagonal AdaGrad; Duchi–Hazan–Singer, JMLR 2011; McMahan–Streeter, COLT 2010). $\mathcal{H} = \{\mathrm{diag}(h) : h \in \mathbb{R}^d_{++}\}$.

$$H_t = \tfrac{1}{\eta}\,\mathrm{diag}\Big(\sqrt{\varepsilon + \textstyle\sum_{s\le t} g_{s,i}^2}\Big)_i, \quad x_{t+1,i} = x_{t,i} - \frac{\eta\, g_{t,i}}{\sqrt{\varepsilon + \sum_{s\le t} g_{s,i}^2}}.$$

One adaptive stepsize per coordinate — most widely deployed form.

John Duchi

Elad Hazan

Yoram Singer

H. B. McMahan

Matthew Streeter

Blockwise AdaGrad-Norm

Example 10.3 (Blockwise AdaGrad-Norm). $d = d_1 + \cdots + d_m$; write $x = (x_1, \ldots, x_m)$, $g_t = (g_{t,1}, \ldots, g_{t,m})$ with $x_\ell, g_{t,\ell} \in \mathbb{R}^{d_\ell}$. $\mathcal{H}_{\text{block}} = \{h_1 I_{d_1} \oplus \cdots \oplus h_m I_{d_m} : h_\ell > 0\}$.

$$H_t = \frac{1}{\eta} \bigoplus_{\ell=1}^{m} \sqrt{\varepsilon + \tfrac{1}{d_\ell} \textstyle\sum_{s \le t} \|g_{s,\ell}\|_2^2}\, I_{d_\ell}, \quad x_{t+1,\ell} = x_{t,\ell} - \frac{\eta}{\sqrt{\varepsilon + \frac{1}{d_\ell} \sum \|g_{s,\ell}\|_2^2}}\, g_{t,\ell}.$$

One scalar per block (commonly one block = one neural-net layer).

Note. "Layerwise AdaGrad" in ML usage often means this *one scalar per layer* form (not diagonal AdaGrad restricted to each layer).

Left-sided Shampoo

Example 10.4 (Left-sided Shampoo; Gupta–Koren–Singer, ICML 2018). $d = d_L d_R$, reshape $x_t = \mathrm{rvec}(X_t)$, $g_t = \mathrm{rvec}(G_t)$; $\mathcal{H}_{\mathrm{left}} = \{H_L \otimes I_{d_R} : H_L \in \mathcal{S}_{++}^{d_L}\}$.

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

Using $(A \otimes I_{d_R})\,\mathrm{rvec}(X) = \mathrm{rvec}(AX)$: $X_{t+1} = X_t - \eta(\varepsilon I_{d_L} + \frac{1}{d_R}\sum G_s G_s^\top)^{-1/2} G_t$.

Left smoothness. $S_{\mathcal{H}_{\mathrm{left}}}(f) =$ smallest $d_R\,\mathrm{tr}(A_L)$ with $\nabla^2 f(X)[\Delta,\Delta] \le \mathrm{tr}(A_L \Delta\Delta^\top)$.

Vineet Gupta

Tomer Koren

Yoram Singer

Block-Diagonal Full AdaGrad

Example 10.5 (Block-diagonal full AdaGrad). $d = d_1 + \cdots + d_m$; $\mathcal{H}_{\mathrm{bd}} = \{H_1 \oplus \cdots \oplus H_m : H_\ell \in \mathcal{S}_{++}^{d_\ell}\}$.

$$H_t = \tfrac{1}{\eta} \bigoplus_{\ell=1}^{m} \Big(\varepsilon I_{d_\ell} + \sum_{s \le t} g_{s,\ell} g_{s,\ell}^\top\Big)^{1/2}, \quad x_{t+1,\ell} = x_{t,\ell} - \eta \Big(\varepsilon I_{d_\ell} + \textstyle\sum_{s \le t} g_{s,\ell} g_{s,\ell}^\top\Big)^{-1/2} g_{t,\ell}.$$

Full adaptive PSD preconditioner inside each block.

Interpolates between blockwise AdaGrad-Norm (Ex 10.3: one scalar per block) and full-matrix AdaGrad (Ex 10.6: dense $d \times d$ preconditioner).

Note. No standard optimizer name in the literature — included to explain the subalgebra hierarchy.

Full-Matrix AdaGrad

Example 10.6 (Full-matrix AdaGrad; Duchi–Hazan–Singer, JMLR 2011; McMahan–Streeter, COLT 2010). $\mathcal{H} = \mathcal{S}^d_{++}$ (all PSD).

$$H_t = \tfrac{1}{\eta}\Big(\varepsilon I_d + \sum_{s\le t} g_s g_s^\top\Big)^{1/2}, \qquad x_{t+1} = x_t - \eta\Big(\varepsilon I_d + \sum_{s\le t} g_s g_s^\top\Big)^{-1/2} g_t.$$

Idealized form; $O(d^3)$ per step via matrix square root.

Selector. $H^{-1} M H^{-1} = \eta^2 I_d \Rightarrow H = \eta^{-1} M^{1/2}$.

John Duchi

Elad Hazan

Yoram Singer

H. B. McMahan

Matthew Streeter

Summary: From Fixed to Adaptive Geometry

Setting	Rate	Ref
Fixed quadratic regret	$\frac{1}{2}\|u - x_1\|_H^2 + \frac{1}{2}\sum \|g_t\|_{H^{-1}}^2$	Cor 10.1
Increasing-metric regret	$2\|X\|_{\mathcal{H}}^2 \operatorname{tr}(H_T) + \frac{1}{2}\sum \|g_t\|_{H_t^{-1}}^2$	Thm 10.3
AdaReg regret	$2\sqrt{2}\,\|X\|_{\mathcal{H}}\big(\|\|\|g_{1:T}\|\|\|_{\mathcal{H}} + d\sqrt{\varepsilon}\big)$	Thm 10.6
Smooth AdaReg	$O\big(\|X\|_{\mathcal{H}}^2 S_{\mathcal{H}}(f)/T\big)$	Thm 10.7

The recipe:

1. Pick a *family* $\mathcal{H}$ of admissible preconditioners (subalgebra).
2. Run AdaReg: maintain $M_t = \varepsilon I + \sum g_s g_s^\top$; choose $H_t = P_{\mathcal{H},\eta}(M_t)$.
3. Quadratic proxy step with H_t.

Matches *best hindsight fixed* $H \in \mathcal{H}$ up to constants. Choice of $\mathcal{H}$ trades off per-step cost vs. rate improvement.

Symbol Review

Geometry		Adaptive selector	
H, H_t	preconditioner	M_t	$\varepsilon I + \sum g_s g_s^\top$ (cumulative)
$\mathcal{H}, \overline{\mathcal{H}}$	family and closed cone	$P_{\mathcal{H},\eta}(M)$	selector minimizer
$\|x\|_H$	$\sqrt{\langle x, Hx\rangle_E}$	$\widehat{P}_{\mathcal{H}}(M)$	normalized ($\mathrm{tr} = 1$)
$\|x\|_{\mathcal{H}}$	family norm	$\Gamma_{\mathcal{H}}$	regularized complexity
$\|X\|_{\mathcal{H}}$	family domain radius	$\|\|\|g_{1:T}\|\|\|_{\mathcal{H}}$	unregularized complexity
Smoothness		**Parameters**	
$\nabla^2 f$	Hessian	η	metric scale parameter
$\mathrm{S}_{\mathcal{H}}(f)$	$\mathcal{H}$-smoothness	ε	regularization of M_0
	$(\inf\{\mathrm{tr}\, A : A \in \mathcal{H}, \nabla^2 f \preceq A\})$	d	$\dim E$