Reading group: Accelerating Variance Reduction for Stochastic Gradient Methods

cvx
smooth

(st) cvx

$$\sum_{i=0}^{n} f_i + g$$

DEPARTMENT OF STATISTICS

UNIVERSITY OF OXFORD

February 22, 2020

Review of Accelerated Gradient Descent.

Aim: Find an ε-minimizer of an L-smooth, differentiable and convex function $f : \mathcal{X} \to \mathbb{R}$ in $\boxed{O(\sqrt{\frac{L}{\varepsilon}})}$ iterations, for convex $\mathcal{X}$. Equivalently, starting at an arbitrary point x_0, for a minimizer x^* and after T iterations, compute a point x_t such that $f(x_t) - f(x^*) \lesssim \boxed{\frac{L}{T^2}} < \varepsilon$.

Reduction: Most of the work is done when going from a 2ε-minimizer to an ε-minimizer. Indeed, assume we can go from $f(x_0) - f(x^*) \leq d$ to $f(x_t) - f(x^*) \leq \frac{d}{2}$ in $t = O(\sqrt{\frac{L}{d}})$. Then we can obtain an ε-minimizer in

$$T \lesssim \boxed{\sqrt{L/d} + \sqrt{L/(d/2)} + \cdots + \sqrt{L/4\varepsilon} + \boxed{\sqrt{L/2\varepsilon}}}$$
$$< \sum_{i=1}^{\infty} \sqrt{L/2^i \varepsilon} \lesssim \sqrt{L/\varepsilon}.$$

Acceleration can be understood as a compromise between Gradient Descent (builds a primal solution) and Mirror Descent (builds a dual solution).

Review of Gradient Descent and Mirror Descent.

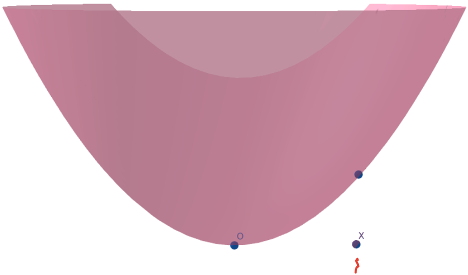


Figure: Parabola $\frac{L}{2}\|x - O\|^2$.

- Assume gradient at x is w.

Review of Gradient Descent and Mirror Descent.

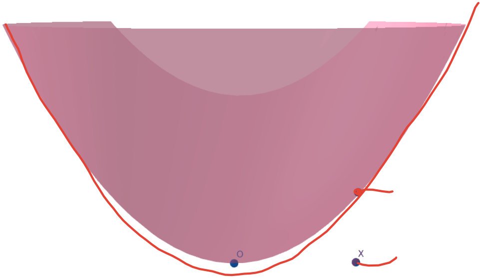


Figure: Parabola $\frac{L}{2}\|x - O\|^2$.

- Assume gradient at x is w.
- What is the gap between the parabola at x and at the optimum?

Review of Gradient Descent and Mirror Descent.

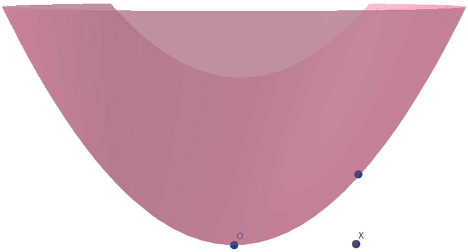


Figure: Parabola $\frac{L}{2}\|x-O\|^2$.

- Assume gradient at x is w.
- What is the gap between the parabola at x and at the optimum?
- We can work in dimension 2.

Review of Gradient Descent and Mirror Descent.

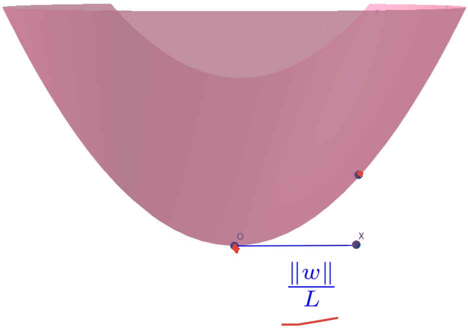


Figure: Parabola $\frac{L}{2}\|x-O\|^2$.

- Assume gradient at x is w.
- What is the gap between the parabola at x and at the optimum?
- We can work in dimension 2.
- $L\|x-O\| = \|w\|$.

$\frac{L\|x-O\|^2}{2} = \frac{\|w\|^2}{2L}$

Review of Gradient Descent and Mirror Descent.

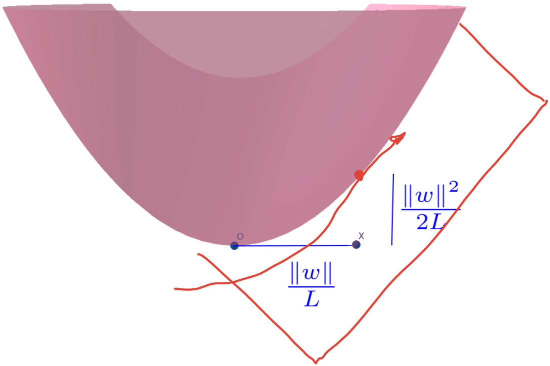


Figure: Parabola $\boxed{\frac{L}{2}}\|x-O\|^2$.

- Assume gradient at x is w.
- What is the gap between the parabola at x and at the optimum?
- We can work in dimension 2.
- $L\|x-O\| = \|w\|$.
- $\frac{L}{2}\|x-O\|^2 = \frac{\|w\|^2}{2L}$.

Review of Gradient Descent and Mirror Descent.

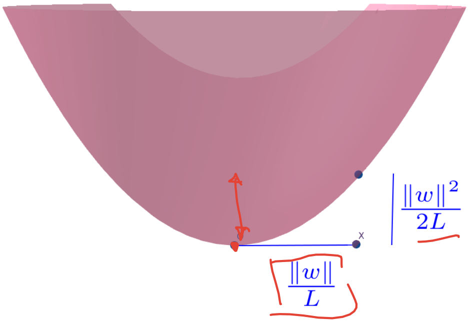


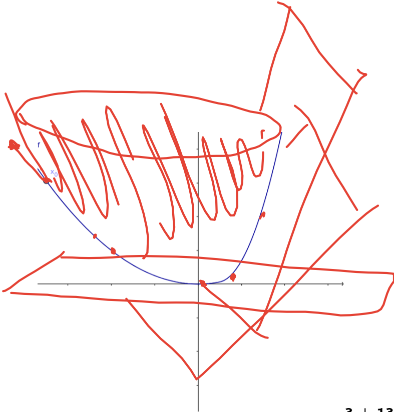
Figure: Parabola $\frac{L}{2}\|x-O\|^2$.

- Assume gradient at x is w.
- What is the gap between the parabola at x and at the optimum?
- We can work in dimension 2.
- $L\|x-O\| = \|w\|$.
- $\frac{L}{2}\|x-O\|^2 = \frac{\|w\|^2}{2L}$.

Gradient descent minimizes the upper bound on the function that smoothness yields. For a gradient $\nabla f(x_t)$ it moves to $x_{t+1} = x_t - \boxed{\frac{1}{L}\nabla f(x_t)}$ to decrease the objective $f(x_t) - f(x_{t+1}) \geq \|\nabla f(x_t)\|^2 / 2L$.

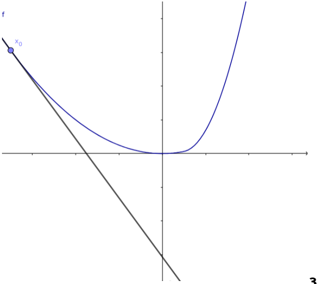
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.



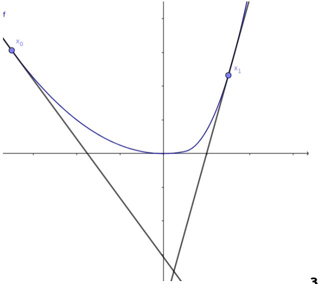
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.



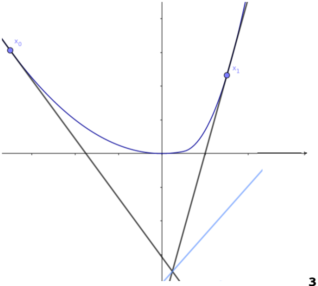
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.



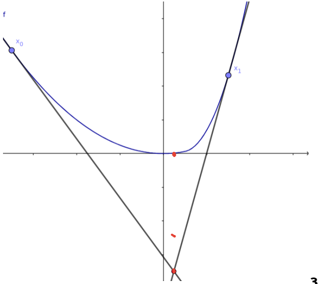
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.



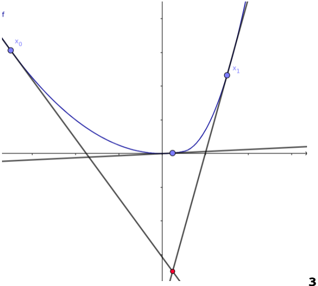
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.
- It follows this principle (approx.): be optimistic and evaluate the function where the lower bound is lowest. If the function value is close to the minimum of the lower bound, we know we are close to finish optimizing. If it is far, we learn by increasing the lower bound. Dual progress!



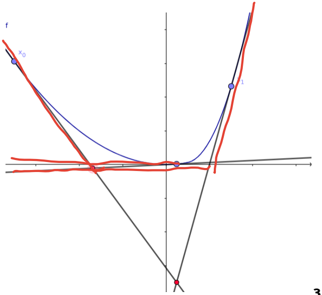
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.
- It follows this principle (approx.): be optimistic and evaluate the function where the lower bound is lowest. If the function value is close to the minimum of the lower bound, we know we are close to finish optimizing. If it is far, we learn by increasing the lower bound. Dual progress!



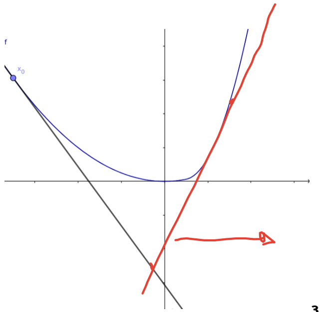
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.
- It follows this principle (approx.): be optimistic and evaluate the function where the lower bound is lowest. If the function value is close to the minimum of the lower bound, we know we are close to finish optimizing. If it is far, we learn by increasing the lower bound. Dual progress!



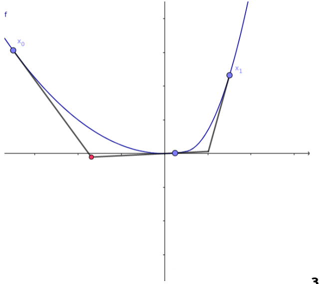
Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.
- It follows this principle (approx.): be optimistic and evaluate the function where the lower bound is lowest. If the function value is close to the minimum of the lower bound, we know we are close to finish optimizing. If it is far, we learn by increasing the lower bound. Dual progress!
- **This intuition needs refinement:** The **lowest point** of the lower bound **could be undefined**. Also, optimizing the natural lower bound, that is the max of all lower bounds, is expensive in time and memory.



Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.
- It follows this principle (approx.): be optimistic and evaluate the function where the lower bound is lowest. If the function value is close to the minimum of the lower bound, we know we are close to finish optimizing. If it is far, we learn by increasing the lower bound. Dual progress!
- **This intuition needs refinement:** The lowest point of the lower bound could be undefined. Also, **optimizing** the natural lower bound, that is the **max of all lower bounds, is expensive** in time and memory.



Review of Mirror Descent. Intuition.

- Convexity yields linear lower bounds.
- Mirror Descent (MD) builds a lower bound with them.
- It follows this principle (approx.): be optimistic and evaluate the function where the lower bound is lowest. If the function value is close to the minimum of the lower bound, we know we are close to finish optimizing. If it is far, we learn by increasing the lower bound. Dual progress!
- **This intuition needs refinement:** The lowest point of the lower bound could be undefined. Also, optimizing the natural lower bound, that is the max of all lower bounds, is expensive in time and memory.

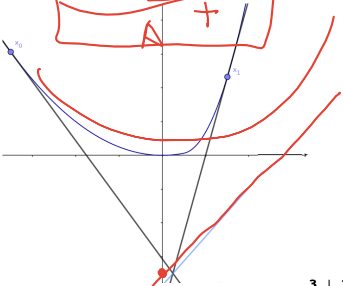
A

$f(\bar{x}) + \sum \langle \nabla f(x_i), x_i - x^* \rangle \le f(x^*) + \frac{\alpha}{2}\|x_{t-1} - x^*\|^2$

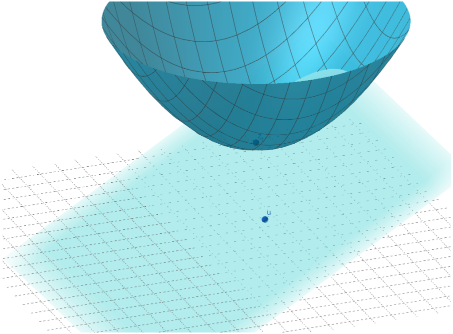
Solutions: Regularize + use average of lower bounds. This is good enough! In equations:

$f(\sum_i x_i/t) - f(x^*)$ (Jensen)
$\le \left(\sum_i f(x_i) - f(x^*)\right)/t$ (convexity)
$\le \left(\sum_i \langle \nabla f(x_i), x_i - x^* \rangle\right)/t$.

If $\alpha \|x_{t-1} - x^*\|^2$ is bounded, adding the regularizer $\alpha \|x_{t-1} - x^*\|^2 / t$ to the lower bound will not change the rate if we aim for $O(1/t)$ or slower.



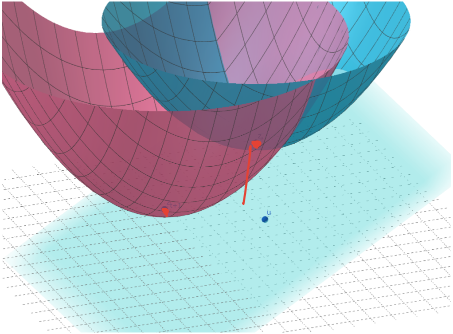
Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2} \left\| \cdot - z_t \right\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t.



Review of Mirror Descent. Mirror Descent Lemma.

Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2}\|\cdot - z_t\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t. Denote z_{t+1} its minimum. The difference of heights must then be $\boxed{\frac{\|w\|^2}{2}}$ (slide 3) and therefore

$p_r(\cdot) = \frac{1}{2}\|\cdot - z_{t+1}\|^2 - \frac{\|w\|^2}{2}$.



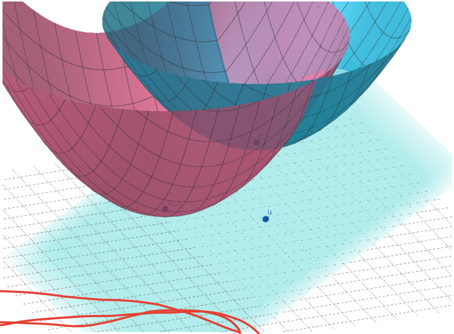
$$p_r = p_b + H.$$

So $\forall u \in \mathbb{R}^d$: $\langle w, z_t - u \rangle = \frac{\|w\|^2}{2} + \frac{1}{2}\|u - z_t\|^2 - \frac{1}{2}\|u - z_{t+1}\|^2$ $(-H = p_b - p_r)$

Review of Mirror Descent. Mirror Descent Lemma.

Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2}\|\cdot - z_t\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t. Denote z_{t+1} its minimum. The difference of heights must then be $\frac{\|w\|^2}{2}$ (slide 3) and therefore

$p_r(\cdot) = \frac{1}{2}\|\cdot - z_{t+1}\|^2 - \frac{\|w\|^2}{2}$.



$$p_r = p_b + H.$$

So $\forall u \in \mathbb{R}^d$: $\langle w, z_t - u\rangle = \frac{\|w\|^2}{2} + \frac{1}{2}\|u - z_t\|^2 - \frac{1}{2}\|u - z_{t+1}\|^2 \quad (-H = p_b - p_r)$

For $w_t = \alpha \nabla f(z_t)$ we converge for L_1-Lipschitz functions (i.e. $\|\nabla f(\cdot)\| \le L_1$):

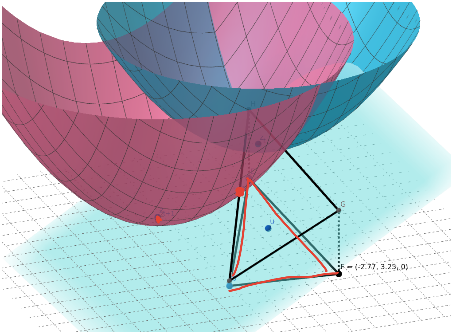
$$\underbrace{f\Big(\sum_{i=0}^{T-1} z_i/T\Big) - f(x^*)}_{O\left(\frac{1}{\sqrt{T}}\right)} \le \frac{\sum_i \langle \alpha\nabla f(z_i), z_i - x^*\rangle}{\alpha T} \le \frac{\alpha^2 L_1^2 T}{2\alpha T} + \frac{1}{2\alpha T}\|x^* - z_0\|^2 - \|x^* - z_T\|$$

$$= \sqrt{L_1^2 \|x^* - z_0\|^2 / 4T}. \text{ (where } \alpha = \sqrt{\|x^* - z_0\|^2 / T L_1^2})$$

Review of Mirror Descent. Mirror Descent Lemma.

Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2} \|\cdot - z_t\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t. Denote z_{t+1} its minimum. The difference of heights must then be $\frac{\|w\|^2}{2}$ (slide 3) and therefore $p_r(\cdot) = \frac{1}{2} \|\cdot - z_{t+1}\|^2 - \frac{\|w\|^2}{2}$.

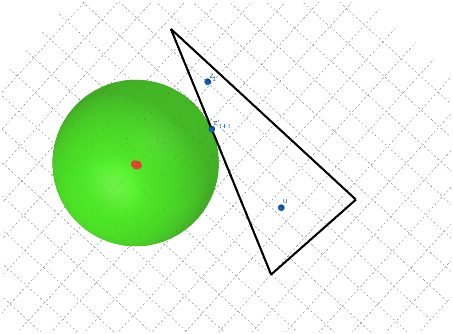
$$p_r = p_b + H.$$



So $\forall u \in \mathbb{R}^d$:$\langle w, z_t - u \rangle = \frac{\|w\|^2}{2} + \frac{1}{2} \|u - z_t\|^2 - \frac{1}{2} \|u - z_{t+1}\|^2$ $(-H = p_b - p_r)$
If we have a constraint $\mathcal{X}$ we need to find the optimum z'_{t+1} of p_r in $\mathcal{X}$.

Review of Mirror Descent. Mirror Descent Lemma.

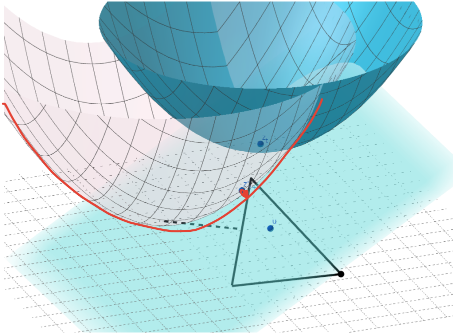
Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2} \|\cdot - z_t\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t. Denote z_{t+1} its minimum. The difference of heights must then be $\frac{\|w\|^2}{2}$ (slide 3) and therefore $p_r(\cdot) = \frac{1}{2} \|\cdot - z_{t+1}\|^2 - \frac{\|w\|^2}{2}$.



$$p_r = p_b + H.$$

So $\forall u \in \mathbb{R}^d$:$\langle w, z_t - u \rangle = \frac{\|w\|^2}{2} + \frac{1}{2} \|u - z_t\|^2 - \frac{1}{2} \|u - z_{t+1}\|^2$ $(-H = p_b - p_r)$
Projecting z_t onto $\mathcal{X}$ we get z_t'. (Minimum dist. to $z_{t+1} \Rightarrow$ min. value of p_r.)

Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2}\|\cdot - z_t\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t. Denote z_{t+1} its minimum. The difference of heights must then be $\frac{\|w\|^2}{2}$ (slide 3) and therefore $p_r(\cdot) = \frac{1}{2}\|\cdot - z_{t+1}\|^2 - \frac{\|w\|^2}{2}$.



$p_r = p_b + H.$

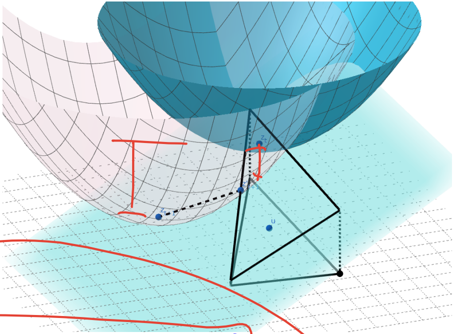
So $\forall u \in \mathbb{R}^d$: $\langle w, z_t - u \rangle = \frac{\|w\|^2}{2} + \frac{1}{2}\|u - z_t\|^2 - \frac{1}{2}\|u - z_{t+1}\|^2$ $(-H = p_b - p_r)$

Here $\forall u \in \mathcal{X}$: $\frac{1}{2}\left\|u - z'_{t+1}\right\|^2 \leq p_r(u) - p_r(z'_{t+1})$ (p_r is strongly convex, z'_{t+1} is minimizer. Alternatively, note p_r grows from z'_{t+1} towards $u \in \mathcal{X}$, Hessian is $\succcurlyeq I$ so difference of p_r's is $\geq \frac{1}{2}\text{dist}^2$ ($\frac{1}{2}\text{dist}^2$ if starting with zero directional derivative and hessian in the path in between evaluates 1 at (v, v) with $v = u - z'_{t+1}$. Higher o/w)).

Review of Mirror Descent. Mirror Descent Lemma.

Given $z_t, w \in \mathbb{R}^d$ we have a hyperplane $H(\cdot) = \langle w, \cdot - z_t \rangle$. We add the blue parabola $p_b(\cdot) = \frac{1}{2}\|\cdot - z_t\|^2$ as a regularizer. The sum is another parabola p_r **tangent** to H at z_t. Denote z_{t+1} its minimum. The difference of heights must then be $\frac{\|w\|^2}{2}$ (slide 3) and therefore

$p_r(\cdot) = \frac{1}{2}\|\cdot - z_{t+1}\|^2 - \frac{\|w\|^2}{2}$.



$p_r = p_b + H.$

So $\forall u \in \mathbb{R}^d$: $\langle w, z_t - u \rangle \leq \frac{\|w\|^2}{2} + \frac{1}{2}\|u - z_t\|^2 - \frac{1}{2}\|u - z_{t+1}\|^2$ $(-H = p_b - p_r)$

Here $\forall u \in \mathcal{X}$: $\frac{1}{2}\|u - z'_{t+1}\|^2 \leq p_r(u) - p_r(z'_{t+1})$ (p_r is strongly convex).

So $\langle w, z_t - u \rangle \leq -p_r(z'_{t+1}) + \frac{1}{2}\|u - z_t\|^2 - \frac{1}{2}\|u - z'_{t+1}\|^2$. Term $-p_r(z'_{t+1})$ can be trivially bounded by $-p_r(z_t) \leq \|w\|^2/2$ (and convergence follows as before) but sometimes it is better to use its exact value

$-p_r(z'_{t+1}) = -p_b(z'_{t+1}) - H(z'_{t+1}) = \langle w, z_t - z'_{t+1} \rangle - \frac{1}{2}\|z_t - z'_{t+1}\|^2$.

Review of Mirror Descent. Mirror Descent Lemma.

So we proved that for all $u \in \mathcal{X} \subseteq \mathbb{R}^d$, $w, z_t \in \mathbb{R}^d$ we have

$$\langle w, z_t - u\rangle \leq \langle w, z_t - z'_{t+1}\rangle - \frac{1}{2}\left\|z_t - z'_{t+1}\right\|^2 + \frac{1}{2}\left\|u - z_t\right\|^2 - \frac{1}{2}\left\|u - z'_{t+1}\right\|^2$$

and in particular

$$\langle w, z_t - u\rangle \leq \frac{\left\|w\right\|^2}{2} + \frac{1}{2}\left\|u - z_t\right\|^2 - \frac{1}{2}\left\|u - z'_{t+1}\right\|^2.$$

Review of Mirror Descent. Mirror Descent Lemma.

So we proved that for all $u \in \mathcal{X} \subseteq \mathbb{R}^d$, $w, z_t \in \mathbb{R}^d$ we have

$$\langle w, z_t - u \rangle \leq \langle w, z_t - z'_{t+1} \rangle - \frac{1}{2} \left\| z_t - z'_{t+1} \right\|^2 + \frac{1}{2} \left\| u - z_t \right\|^2 - \frac{1}{2} \left\| u - z'_{t+1} \right\|^2$$

and in particular

$$\langle w, z_t - u \rangle \leq \frac{\|w\|^2}{2} + \frac{1}{2} \left\| u - z_t \right\|^2 - \frac{1}{2} \left\| u - z'_{t+1} \right\|^2 .$$

When w is a multiple of $\nabla f(z_t)$ (what you would normally want) the term $\|w\|^2 / 2$ can be large sometimes. But recall, if we use gradient descent from z_t, the new point y_{t+1} has a guaranteed progress proportional to $\|w\|^2$. Good GD progress means compensating bad MD performance and bad GD progress would happen only when MD has good performance!! **Problem:** each method will tell us to evaluate a different point . We need to mix them in some way: **Linear coupling**.

Linear Coupling (Unconstrained)

Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = (1-\tau)y_t + \tau z_t'$, where y_t and z_t' are the gradient and mirror points defined as $y_0 = z_0' = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y \in \mathcal{X}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 \right\}$ and $z_t' = \operatorname{argmin}_{z \in \mathcal{X}} \left\{ \langle \alpha \nabla f(x_t), z - x_t \rangle + \frac{1}{2} \|z - z_t'\|^2 \right\}$, for α to be chosen later.

Linear Coupling (Unconstrained)

Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = (1-\tau)y_t + \tau z'_t$, where y_t and z'_t are the gradient and mirror points defined as $y_0 = z'_0 = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y \in \mathcal{X}}\{f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2}\|y - x_t\|^2\}$ and $z'_t = \operatorname{argmin}_{z \in \mathcal{X}}\{\langle \alpha \nabla f(x_t), z - x_t \rangle + \frac{1}{2}\|z - z'_t\|^2\}$, for α to be chosen later.

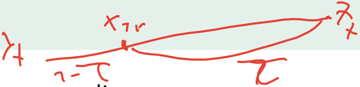
$x_{t+1} - z_t$

Analysis: $f(\underline{\sum_{t=0}^{T-1} x_t/T}) - f(x^*) \le \underline{\frac{1}{T}\sum_{t=0}^{T-1}\langle \nabla f(x_{t+1}), x_{t+1} - \underline{x^* \pm z'_t}\rangle \le} (*)$

Linear Coupling (Unconstrained)

Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = \boxed{(1-\tau)y_t + \tau z'_t}$, where y_t and z'_t are the gradient and mirror points defined as $y_0 = z'_0 = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y \in \mathcal{X}} \{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 \}$ and $z'_t = \operatorname{argmin}_{z \in \mathcal{X}} \{ \langle \alpha \nabla f(x_t), z - x_t \rangle + \frac{1}{2} \|z - z'_t\|^2 \}$, for α to be chosen later.
Analysis: $f(\sum_{t=0}^{T-1} x_t / T) - f(x^*) \leq \frac{1}{T} \sum_{t=0}^{T-1} \langle \nabla f(x_{t+1}), x_{t+1} - x^* \pm z'_t \rangle \leq (*)$
Parameter τ is chosen to satisfy $\boxed{(1-\tau)/\tau = \alpha L}$ to balance the constant between the progress of GD and $\|w\|^2$.

Linear Coupling (Unconstrained)



Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = (1-\tau)y_t + \tau z_t'$, where y_t and z_t' are the gradient and mirror points defined as $y_0 = z_0' = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y \in \mathcal{X}}\{f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2}\|y - x_t\|^2\}$ and $z_t' = \operatorname{argmin}_{z \in \mathcal{X}}\{\langle \alpha \nabla f(x_t), z - x_t \rangle + \frac{1}{2}\|z - z_t'\|^2\}$, for α to be chosen later.
Analysis: $f(\sum_{t=0}^{T-1} x_t/T) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}\langle \nabla f(x_{t+1}), x_{t+1} - x^* \pm z_t' \rangle \leq (*)$
Parameter τ is chosen to satisfy $(1-\tau)/\tau = \alpha L$ to balance the constant between the progress of GD and $\|w\|^2$. That is, the coupling implies

$$\langle \nabla f(x_{t+1}), x_{t+1} - z_t' \rangle = \frac{1-\tau}{\tau}\langle \nabla f(x_{t+1}), y_t - x_t \rangle \overset{\text{cvx.}}{\leq} \alpha L(f(y_t) - f(x_{t+1})) =: A_t.$$

Linear Coupling (Unconstrained)

Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = (1-\tau)y_t + \tau z'_t$, where y_t and z'_t are the gradient and mirror points defined as $y_0 = z'_0 = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y \in \mathcal{X}}\{f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2\}$ and $z'_t = \operatorname{argmin}_{z \in \mathcal{X}}\{\langle \alpha \nabla f(x_t), z - x_t \rangle + \frac{1}{2} \|z - z'_t\|^2\}$, for α to be chosen later.
Analysis: $f(\sum_{t=0}^{T-1} x_t/T) - f(x^*) \leq \frac{1}{T} \sum_{t=0}^{T-1} \langle \nabla f(x_{t+1}), x_{t+1} - x^* \pm z'_t \rangle \leq (*)$
Parameter τ is chosen to satisfy $(1-\tau)/\tau = \alpha L$ to balance the constant between the progress of GD and $\|w\|^2$. That is, the coupling implies
$\langle \nabla f(x_{t+1}), x_{t+1} - z'_t \rangle = \frac{1-\tau}{\tau} \langle \nabla f(x_{t+1}), y_t - x_t \rangle \overset{\text{cvx.}}{\leq} \alpha L (f(y_t) - f(x_{t+1})) =: A_t$.
And finally, picking $w_t = \alpha \nabla f(x_t)$, and using the unconstrained GD guarantee $\alpha \|\nabla f(x_{t+1})\|^2 /2 \leq \alpha L (f(x_{t+1}) - f(y_{t+1})) =: B_t$, we obtain:

Linear Coupling (Unconstrained)

Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = (1-\tau)y_t + \tau z_t'$, where y_t and z_t' are the gradient and mirror points defined as $y_0 = z_0' = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y\in\mathcal{X}}\{f(x_t) + \langle\nabla f(x_t), y - x_t\rangle + \frac{L}{2}\|y - x_t\|^2\}$ and $z_t' = \operatorname{argmin}_{z\in\mathcal{X}}\{\langle\alpha\nabla f(x_t), z - x_t\rangle + \frac{1}{2}\|z - z_t'\|^2\}$, for α to be chosen later.
Analysis: $f(\sum_{t=0}^{T-1} x_t/T) - f(x^*) \le \frac{1}{T}\boxed{\sum_{t=0}^{T-1}}\langle\nabla f(x_{t+1}), x_{t+1} - x^* \pm z_t'\rangle \le (*)$
Parameter τ is chosen to satisfy $(1-\tau)/\tau = \alpha L$ to balance the constant between the progress of GD and $\|w\|^2$. That is, the coupling implies

$$\boxed{\langle\nabla f(x_{t+1}), x_{t+1} - z_t'\rangle} = \frac{1-\tau}{\tau}\langle\nabla f(x_{t+1}), y_t - x_t\rangle \overset{\text{cvx.}}{\le} \alpha L(f(y_t) - f(x_{t+1})) =: A_t.$$

And finally, picking $w_t = \alpha\nabla f(x_t)$, and using the unconstrained GD guarantee $\alpha\|\nabla f(x_{t+1})\|^2/2 \le \boxed{\alpha L(f(x_{t+1})} - f(y_{t+1})) =: B_t$, we obtain:

$z_t - x^*$

$$\langle\alpha\nabla f(x_{t+1}), x_{t+1} - x^* \pm z_t'\rangle/\alpha \le \frac{\alpha\|\nabla f(x_{t+1})\|^2}{2} + \underbrace{\frac{\|x^* - z_t'\|^2 - \|x^* - z_{t+1}'\|^2}{2\alpha}}_{=:C_t} + A_t$$

$$\le A_t + B_t + C_t = \alpha L(f(y_t) - f(y_{t+1})) + C_t.$$

Linear Coupling (Unconstrained)

Linear coupling: Run your MD, but compute the next gradient at a convex (linear) combination of the points suggested $x_{t+1} = (1-\tau)y_t + \tau z'_t$, where y_t and z'_t are the gradient and mirror points defined as $y_0 = z'_0 = x_0 \in \mathcal{X}$ arbitrary, $y_t = \operatorname{argmin}_{y\in\mathcal{X}}\{f(x_t) + \langle \nabla f(x_t), y - x_t\rangle + \frac{L}{2}\|y - x_t\|^2\}$ and $z'_t = \operatorname{argmin}_{z\in\mathcal{X}}\{\langle \alpha\nabla f(x_t), z - x_t\rangle + \frac{1}{2}\|z - z'_t\|^2\}$, for α to be chosen later.

Analysis: $f(\sum_{t=0}^{T-1} x_t/T) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}\langle \nabla f(x_{t+1}), x_{t+1} - x^* \pm z'_t\rangle \leq (*)$

Parameter τ is chosen to satisfy $(1-\tau)/\tau = \alpha L$ to balance the constant between the progress of GD and $\|w\|^2$. That is, the coupling implies

$\langle \nabla f(x_{t+1}), x_{t+1} - z'_t\rangle = \frac{1-\tau}{\tau}\langle \nabla f(x_{t+1}), y_t - x_t\rangle \overset{\text{cvx.}}{\leq} \alpha L(f(y_t) - f(x_{t+1})) =: A_t$.

And finally, picking $w_t = \alpha\nabla f(x_t)$, and using the unconstrained GD guarantee $\alpha\|\nabla f(x_{t+1})\|^2/2 \leq \alpha L(f(x_{t+1}) - f(y_{t+1})) =: B_t$, we obtain:

$$\langle \alpha\nabla f(x_{t+1}), x_{t+1} - x^* \pm z'_t\rangle/\alpha \leq \frac{\alpha\|\nabla f(x_{t+1})\|^2}{2} + \underbrace{\frac{\|x^* - z'_t\|^2 - \|x^* - z'_{t+1}\|^2}{2\alpha}}_{=:C_t} + A_t$$

$$\leq A_t + B_t + C_t = \alpha L(f(y_t) - f(y_{t+1})) + C_t.$$

$(*) \leq \frac{1}{T}\left(\alpha L(f(y_0) - f(y_T)) + \frac{1}{2\alpha}\|x^* - z_0\|^2\right) \leq \frac{\sqrt{dL\|x^*-x_0\|^2}}{T} \leq \frac{d}{2}$, for $T = O((\frac{L}{d})^{\frac{1}{2}})$ and $\alpha = (\frac{\|x^*-x_0\|^2}{2Ld})^{\frac{1}{2}}$. The reduction applies.

Linear Coupling (Constrained)

Mirror lemma. Define v such that $(v - x_{k+1})/\tau = (z_t - z_{t+1})$, for $\tau = \overset{1}{\alpha L}$.

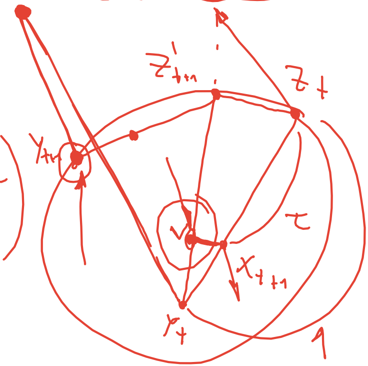
$$\underline{\langle w, z_t - u \rangle} \le \boxed{\langle w, z_t - z'_{t+1} \rangle - \frac{1}{2} \left\| z_t - z'_{t+1} \right\|^2} + \frac{1}{2} \left\| u - z_t \right\|^2 - \frac{1}{2} \left\| u - z'_{t+1} \right\|^2$$

$\wedge$

$$\alpha L \langle \alpha \nabla f, v - x_{k+1} \rangle - \frac{\alpha^2 L^2}{2} \| v - x_{k+1} \|^2$$

$$\alpha^2 L \left(f(x_{k+1}) + \langle \nabla f, v - x_{k+1} \rangle - \frac{L}{2} \| v - x_{k+1} \|^2 \right)$$

$$\le \alpha^2 L \, \mathrm{Progress}(x_{k+1})$$



Finite sum stochastic convex optimization

Problem:

$$O\left(\frac{nL}{T^2}\right)$$

$$\min_{x\in\mathbb{R}^m}\left\{F(x) \overset{\text{def}}{=} f(x)+g(x) \overset{\text{def}}{=} \boxed{\frac{1}{n}\sum_{i=1}^{n} f_i(x)} + g(x)\right\}. \quad O\left(\frac{nL}{T}\right)$$

For differentiable L-smooth convex $f_i : \mathbb{R}^m \to \mathbb{R}$ and proper, μ-strongly convex $(\mu \geq 0)$ $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ that can be non-differentiable but it is lower semicontinuous.

$$\operatorname{prox}_g(y) = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|x-y\|^2 + g(x)\right\}$$

MSEB property: For $\rho_B, \rho_F, \rho_M \in (0,1]$:

$$\nabla f\left(x_{k+1}\right) - \mathbb{E}_k\widetilde{\nabla}_{k+1} = \left(1-\rho_B\right)\left(\nabla f\left(x_k\right) - \widetilde{\nabla}_k\right) \quad \leftarrow \text{bias}$$

$$\mathbb{E}\left\|\widetilde{\nabla}_{k+1} - \nabla f\left(x_{k+1}\right)\right\|^2 \leq \mathcal{M}_k \quad \leftarrow \text{MSE}$$

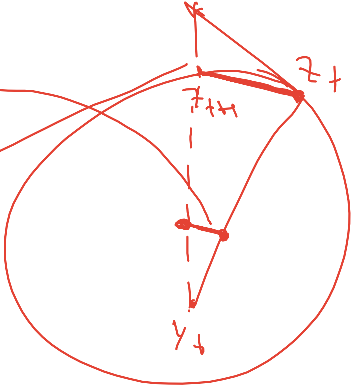
$$\mathcal{M}_k \leq \frac{M_1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right)\right\|^2 + \mathcal{F}_k + \left(1-\rho_M\right)\mathcal{M}_{k-1}$$

$$\mathcal{F}_k \leq \sum_{\ell=0}^{k}\frac{M_2\left(1-\rho_F\right)^{k-\ell}}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right)\right\|^2$$

Algorithm

Given an estimator of the gradient $\tilde{\nabla}_{k+1}$The algorithm is the following, with proper learning rates γ_k and linear coupling parameters τ_k.

Initialize $x_0 = y_0 = z_0$.
for $k = 0, 1, \ldots, T-1$ **do**
    $x_{k+1} \leftarrow \tau_k z_k + (1 - \tau_k) y_k$.
    Compute $\widetilde{\nabla}_{k+1}$, an estimate of $\nabla f(x_{k+1})$.
    $z_{k+1} \leftarrow \mathrm{prox}_{\gamma_k g}\left(z_k - \gamma_k \widetilde{\nabla}_{k+1}\right)$.
    $y_{k+1} \leftarrow \tau_k z_{k+1} + (1 - \tau_k)\, y_k$
end for
return y_t.



Finite sum stochastic convex optimization

Common Estimators: (authors explicit rates for these + SARGE)

SVRG:

$\tilde{x}$

$$\widetilde{\nabla}_{k+1}^{\mathsf{SVRG}} \overset{\mathrm{def}}{=} \frac{1}{|B_k|}\left(\sum_{b_j\in B_k} \nabla f_{b_j}\left(x_{k+1}\right) - \nabla f_{b_j}(\widetilde{x})\right) + \nabla f(\widetilde{x})$$

x_{k+1}

$O(n)$

SAGA:

$$\widetilde{\nabla}_{k+1}^{\mathsf{SAGA}} \overset{\mathrm{def}}{=} \frac{1}{|B_k|}\left(\sum_{b_j\in B_k} \nabla f_{b_j}\left(x_{k+1}\right) - \nabla f_{b_j}\left(\varphi_k^{b_j}\right)\right) + \boxed{\frac{1}{n}\sum_{i=1}^{n} \nabla f_i\left(\varphi_k^{i}\right)}$$

$\varphi_{k+1}^{b_j} = \nabla f_{b_j}(x_{k+1})$

SARAH:

$$\widetilde{\nabla}_{k+1}^{\mathsf{SARAH}} \overset{\mathrm{def}}{=} \begin{cases} \frac{1}{|B_k|}\left(\sum_{b_j\in B_k} \nabla f_{b_j}\left(x_{k+1}\right) - \nabla f_{b_j}\left(x_k\right)\right) + \widetilde{\nabla}_k^{\mathsf{SARAH}} & \text{w.p. } 1-\frac{1}{p}, \\ \nabla f\left(x_{k+1}\right) & \text{w.p. } \frac{1}{p} \end{cases}$$

Bound of Main Result

$$\gamma_k \left(f\left(x_{k+1}\right) - f\left(x^*\right)\right)$$

Convexity.

Bound of Main Result

$$\gamma_k(f(x_{k+1}) - f(x^*))$$

$$\leq \gamma_k \langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle$$

Convexity.

Bound of Main Result

$$= \gamma_k \langle \nabla f\left(x_{k+1}\right), x_{k+1} - x^* \rangle$$

Add and substract z_k. Split.

Bound of Main Result

$$\gamma_k \langle \nabla f\left(x_{k+1}\right), x_{k+1} - z_k \rangle + \gamma_k \langle \nabla f\left(x_{k+1}\right), z_k - x^* \rangle$$

$$= \gamma_k \langle \nabla f\left(x_{k+1}\right), x_{k+1} - x^* \rangle$$

Add and substract z_k. Split.

Bound of Main Result

$$\gamma_k \langle \nabla f(x_{k+1}), x_{k+1} - z_k \rangle + \gamma_k \langle \nabla f(x_{k+1}), z_k - x^* \rangle$$

Linear coupling: $\boxed{x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k}$

Bound of Main Result

$$\gamma_k \langle \nabla f(x_{k+1}), x_{k+1} - z_k \rangle + \gamma_k \langle \nabla f(x_{k+1}), z_k - x^* \rangle$$

$$= (\gamma_k (1 - \tau_k) / \tau_k) \langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle + \gamma_k \langle \nabla f(x_{k+1}), z_k - x^* \rangle$$

Linear coupling: $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$

Bound of Main Result

$$= (\gamma_k (1 - \tau_k) / \tau_k) \langle \nabla f (x_{k+1}), y_k - x_{k+1} \rangle + \gamma_k \langle \nabla f (x_{k+1}), z_k - x^* \rangle$$

Make $D_f(y_k, x_{k+1})$ appear from the first term. Make $\tilde{\nabla}_{k+1}$ appear in the second term.

Bound of Main Result

$$D_f(y,x) = f(y) - f(x) - \boxed{\langle \nabla f(x), y - x \rangle}$$

$$\gamma_k(1-\tau_k)/\tau_k(f(y_k) - f(x_{k+1})) - \gamma_k(1-\tau_k)/\tau_k D_f(y_k, x_{k+1})$$

$$= (\gamma_k(1-\tau_k)/\tau_k)\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \gamma_k\langle \nabla f(x_{k+1}), z_k - x^*\rangle$$

$$+\gamma_k\langle \widetilde{\nabla}_{k+1}, z_k - x^*\rangle + \gamma_k\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\rangle$$

Make $D_f(y_k, x_{k+1})$ appear from the first term. Make $\tilde{\nabla}_{k+1}$ appear in the second term.

Bound of Main Result

$$\gamma_k \left(1-\tau_k\right)/\tau_k \left(f\left(y_k\right)-f\left(x_{k+1}\right)\right)-\gamma_k \left(1-\tau_k\right)/\tau_k D_f \left(y_k, x_{k+1}\right)$$

$$+\gamma_k \langle \widetilde{\nabla}_{k+1}, z_k - x^* \rangle + \gamma_k \langle \nabla f\left(x_{k+1}\right) - \widetilde{\nabla}_{k+1}, z_k - x^* \rangle$$

We add and substract $\boxed{z_{k+1}}$ to ease the comparison.

Bound of Main Result

$$\gamma_k\left(1-\tau_k\right)/\tau_k\left(f\left(y_k\right)-f\left(x_{k+1}\right)\right)-\gamma_k\left(1-\tau_k\right)/\tau_k D_f\left(y_k,x_{k+1}\right)$$

$$\gamma_k\left(1-\tau_k\right)/\tau_k\left(f\left(y_k\right)-f\left(x_{k+1}\right)\right)-\gamma_k\left(1-\tau_k\right)/\tau_k D_f\left(y_k,x_{k+1}\right)$$

$$+\gamma_k\langle\widetilde{\nabla}_{k+1},z_k-x^*\rangle+\gamma_k\langle\nabla f\left(x_{k+1}\right)-\widetilde{\nabla}_{k+1},z_k-x^*\rangle$$

$$+\gamma_k\langle\widetilde{\nabla}_{k+1},z_k-z_{k+1}\rangle+\gamma_k\langle\widetilde{\nabla}_{k+1},z_{k+1}-x^*\rangle$$

$$+\gamma_k\langle\nabla f\left(x_{k+1}\right)-\widetilde{\nabla}_{k+1},z_k-x^*\rangle$$

We add and substract z_{k+1} to ease the comparison.

Bound of Main Result

$$\gamma_k\left(1-\tau_k\right)/\tau_k\left(f\left(y_k\right)-f\left(x_{k+1}\right)\right)-\gamma_k\left(1-\tau_k\right)/\tau_k D_f\left(y_k,x_{k+1}\right)$$

$$+\gamma_k\langle\widetilde{\nabla}_{k+1},z_k-z_{k+1}\rangle+\gamma_k\langle\widetilde{\nabla}_{k+1},z_{k+1}-x^*\rangle$$

$$+\gamma_k\langle\nabla f\left(x_{k+1}\right)-\widetilde{\nabla}_{k+1},z_k-x^*\rangle$$

Algorithm: $x_{k+1}-\underline{y_{k+1}}=\tau_k(z_k-z_{k+1})$

Bound of Main Result

$$\gamma_k\left(1-\tau_k\right)/\tau_k\left(f\left(y_k\right)-f\left(x_{k+1}\right)\right)-\gamma_k\left(1-\tau_k\right)/\tau_k D_f\left(y_k,x_{k+1}\right)$$

$$\gamma_k\left(1-\tau_k\right)/\tau_k\left(f\left(y_k\right)-f\left(x_{k+1}\right)\right)-\gamma_k\left(1-\tau_k\right)/\tau_k D_f\left(y_k,x_{k+1}\right)$$

$$+\gamma_k/\tau_k\langle\widetilde{\nabla}_{k+1},x_{k+1}-y_{k+1}\rangle+\gamma_k\langle\widetilde{\nabla}_{k+1},z_k-x^*\rangle$$

$$+\gamma_k\langle\widetilde{\nabla}_{k+1},z_k-z_{k+1}\rangle+\gamma_k\langle\widetilde{\nabla}_{k+1},z_{k+1}-x^*\rangle$$

$$+\gamma_k\langle\nabla f\left(x_{k+1}\right)-\widetilde{\nabla}_{k+1},z_{k+1}-x^*\rangle$$

$$+\gamma_k\langle\nabla f\left(x_{k+1}\right)-\widetilde{\nabla}_{k+1},z_k-x^*\rangle$$

Algorithm: $x_{k+1}-y_{k+1}=\tau_k(z_k-z_{k+1})$

Bound of Main Result

$$\gamma_k (1-\tau_k)/\tau_k (f(y_k) - f(x_{k+1})) - \gamma_k (1-\tau_k)/\tau_k D_f (y_k, x_{k+1})$$

$$+ \boxed{\gamma_k/\tau_k \langle \widetilde{\nabla}_{k+1}, x_{k+1} - y_{k+1} \rangle} + \boxed{\gamma_k \langle \widetilde{\nabla}_{k+1}, z_k - x^* \rangle}$$

$$+ \gamma_k \langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_{k+1} - x^* \rangle$$

We will bound these two terms separately.

Bound of Main Result

$$\boxed{\frac{\gamma_k}{\tau_k}\langle\widetilde{\nabla}_{k+1}, x_{k+1} - y_{k+1}\rangle}$$

Compare to $\nabla f(x_{k+1})$.

Bound of Main Result

$$\boxed{\frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1}, x_{k+1} - y_{k+1}\rangle}$$

$$\frac{\gamma_k}{\tau_k}\langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle + \frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle$$

Compare to $\nabla f(x_{k+1})$.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}\langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle + \frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle$$

We use smoothness.

Bound of Main Result

$$\leq \frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1}) + (L/2)\left\|x_{k+1} - y_{k+1}\right\|^2\right)$$

$$\frac{\gamma_k}{\tau_k}\langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle + \frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle$$

$$+\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle$$

We use smoothness.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1}) + (L/2)\left\|x_{k+1} - y_{k+1}\right\|^2\right)$$

$$+\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle$$

And Young's inequality.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1}) + (L/2)\left\|x_{k+1} - y_{k+1}\right\|^2\right)$$

$$\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1}) + (L/2)\left\|x_{k+1} - y_{k+1}\right\|^2\right)$$

$$+\frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle$$

$$\leq \gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2 + \frac{1}{4\tau_k^2}\left\|x_{k+1} - y_{k+1}\right\|^2$$

And Young's inequality.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1}) + (L/2)\left\|x_{k+1} - y_{k+1}\right\|^2\right)$$

$$\gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2 + \frac{1}{4\tau_k^2}\left\|x_{k+1} - y_{k+1}\right\|^2$$

Group terms and make F appear.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - f(y_{k+1}) + \frac{\gamma_k}{\tau_k}(f(x_{k+1}) - \boxed{F(y_{k+1}) + g(y_{k+1})})$$

$$\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - f(y_{k+1}) + (L/2)\left\|x_{k+1} - y_{k+1}\right\|^2)$$

$$\leq \underbrace{\left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)}\left\|x_{k+1} - y_{k+1}\right\|^2 + \gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2$$

$$\gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2 + \frac{1}{4\tau_k^2}\left\|x_{k+1} - y_{k+1}\right\|^2$$

Group terms and make F appear.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - f(y_{k+1}) + \frac{\gamma_k}{\tau_k}(f(x_{k+1}) - F(y_{k+1}) + g(y_{k+1}))$$

$$\left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 + \gamma_k^2 \left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2$$

Algorithm: $y_{k+1} = \tau_k z_{k+1} + (1-\tau_k)y_k$. Convexity of g.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - f(y_{k+1}) + \frac{\gamma_k}{\tau_k}(f(x_{k+1}) - F(y_{k+1}) + g(y_{k+1}))$$

$$\leq \frac{\gamma_k}{\tau_k}(f(x_{k+1}) - F(y_{k+1})) + \gamma_k g\left(z_{k+1}\right) + \frac{\gamma_k\left(1-\tau_k\right)}{\tau_k} g\left(y_k\right)$$

$$\left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 + \gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2$$

$$+\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - f(y_{k+1})$$

$$\left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 + \gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2$$

Algorithm: $y_{k+1} = \tau_k z_{k+1} + (1-\tau_k) y_k$. Convexity of g.

Bound of Main Result

$$\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - F(y_{k+1})) + \gamma_k g\left(z_{k+1}\right) + \frac{\gamma_k\left(1-\tau_k\right)}{\tau_k} g\left(y_k\right)$$

$$+\frac{\gamma_k}{\tau_k}(f(x_{k+1}) - f(y_{k+1})$$

$$\left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\left\|x_{k+1} - y_{k+1}\right\|^2 + \gamma_k^2\left\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\right\|^2$$

Now we bound the other term.

Bound of Main Result

$$\boxed{\gamma_k \langle \widetilde{\nabla}_{k+1}, z_k - x^* \rangle}$$

Proximal Lemma.

Bound of Main Result

$$\boxed{\gamma_k \langle \widetilde{\nabla}_{k+1}, z_k - x^* \rangle}$$

$$\leq \frac{1}{2} \left\| z_k - x^* \right\|^2 - \frac{1 + \mu\gamma_k}{2} \left\| z_{k+1} - x^* \right\|^2 - \frac{1}{2} \left\| z_{k+1} - z_k \right\|^2$$

$$-\gamma_k g\left(z_{k+1}\right) + \gamma_k g\left(x^*\right)$$

Proximal Lemma.

Bound of Main Result

$$\frac{1}{2}\left\|z_k - x^*\right\|^2 - \frac{1+\mu\gamma_k}{2}\left\|z_{k+1} - x^*\right\|^2 - \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2$$

$$-\gamma_k g\left(z_{k+1}\right) + \gamma_k g\left(x^*\right)$$

Algorithm: $x_{k+1} - y_{k+1} = \tau_k(z_{k+1} - z_k)$.

Bound of Main Result

$$\leq \frac{1}{2}\left\|z_k - x^*\right\|^2 - \frac{1+\mu\gamma_k}{2}\left\|z_{k+1} - x^*\right\|^2 - \frac{1}{2\tau_k^2}\left\|x_{k+1} - y_{k+1}\right\|^2$$

$$\frac{1}{2}\left\|z_k - x^*\right\|^2 - \frac{1+\mu\gamma_k}{2}\left\|z_{k+1} - x^*\right\|^2 - \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2$$

$$-\gamma_k g\left(z_{k+1}\right) + \gamma_k g\left(x^*\right)$$

$$-\gamma_k g\left(z_{k+1}\right) + \gamma_k g\left(x^*\right)$$

Bound of Main Result

$$\frac{1}{2}\left\|z_k - x^*\right\|^2 - \frac{1+\mu\gamma_k}{2}\left\|z_{k+1} - x^*\right\|^2 - \frac{1}{2\tau_k^2}\left\|x_{k+1} - y_{k+1}\right\|^2$$

$$-\gamma_k g\left(z_{k+1}\right) + \gamma_k g\left(x^*\right)$$

Going back to the original inequality and adding all up we have:

Bound of Main Result

$$
\begin{aligned}
F(y_k) - F(x^*) \overset{H_k}{\leq} & \boxed{\frac{1}{\tau_k} F(y_k) - \frac{1}{\tau_k} F(y_{k+1})} \\
& + \frac{1}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k \gamma_k}\right) \|x_{k+1} - y_{k+1}\|^2 + \boxed{\frac{1}{2\gamma_k}\|z_k - x^*\|^2} \\
& \boxed{- \frac{1+\mu\gamma_k}{2\gamma_k}\|z_{k+1} - x^*\|^2} + \boxed{\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^* \right\rangle} \\
& - \frac{(1-\tau_k)}{\tau_k} D_f(y_k, x_{k+1}) + \boxed{\gamma_k \left\| \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}) \right\|^2}
\end{aligned}
$$

bias

MSE

$$\tau = \frac{1}{\alpha L}$$

We will add up $\sum_{k=0}^{T-1} \gamma_k H_k$.

Bound on MSE

Remember the bound on the MSE

$$\mathbb{E}\left\|\widetilde{\nabla}_{k+1} - \nabla f\left(x_{k+1}\right)\right\|^2 \leq \mathcal{M}_k$$

$$\mathcal{M}_k \leq \frac{M_1}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right)\right\|^2 + \mathcal{F}_k + \left(1 - \rho_M\right) \mathcal{M}_{k-1}$$

$$\mathcal{F}_k \leq \sum_{\ell=0}^{k} \frac{M_2\left(1 - \rho_F\right)^{k-\ell}}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right)\right\|^2$$

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^T \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} s_k^2 \mathcal{M}_k \leq \sum_{k=0}^{T-1} s_k^2 \mathbb{E} \left\| \nabla f\left(x_{k+1}\right) - \widetilde{\nabla}_{k+1} \right\|^2$$

Use the definition of MSE.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq 0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} s_k^2 \mathcal{M}_k \leq \sum_{k=0}^{T-1} s_k^2 \mathbb{E} \left\| \nabla f\left(x_{k+1}\right) - \widetilde{\nabla}_{k+1} \right\|^2$$

$$\leq \sum_{k=0}^{T-1} \frac{M_1 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right) \right\|^2 + s_k^2 \mathcal{F}_k + s_k^2\left(1-\rho_M\right) \mathcal{M}_{k-1}$$

Use the definition of MSE.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \frac{M_1 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right)\right\|^2 + s_k^2 \mathcal{F}_k + s_k^2\left(1-\rho_M\right) \mathcal{M}_{k-1}$$

Substitute the value of $\mathcal{F}_k$. Use $\sum_{k=0}^{j}(1-\rho_F)^k \leq \frac{1}{\rho_F}$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq 0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\leq \sum_{k=0}^{T-1} \frac{\left(M_1\rho_F + 2M_2\right) s_k^2}{n\rho_F} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right) \right\|^2 + s_k^2 \left(1 - \rho_M\right) \mathcal{M}_{k-1}$$

$$\sum_{k=0}^{T-1} \frac{M_1 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right) \right\|^2 + s_k^2 \mathcal{F}_k + s_k^2 \left(1 - \rho_M\right) \mathcal{M}_{k-1}$$

Substitute the value of $\mathcal{F}_k$. Use $\sum_{k=0}^{j}(1-\rho_F)^k \leq \frac{1}{\rho_F}$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \frac{(M_1\rho_F + 2M_2)\, s_k^2}{n\rho_F} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right) \right\|^2 + s_k^2 \left(1 - \rho_M\right) \mathcal{M}_{k-1}$$

Use definition of Θ_2. Recur the inequality over $\mathcal{M}_{k-1}, \mathcal{M}_{k-2}, \ldots$

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \frac{(M_1\rho_F + 2M_2)\, s_k^2}{n\rho_F} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\right\|^2 + s_k^2(1-\rho_M)\,\mathcal{M}_{k-1}$$

$$\leq \sum_{k=0}^{T-1} \sum_{\ell=1}^{k} \frac{\Theta_2 s_k^2 (1-\rho_M)^{k-\ell} \rho_M}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i(x_{\ell+1}) - \nabla f_i(x_\ell)\right\|^2$$

Use definition of Θ_2. Recur the inequality over $\mathcal{M}_{k-1}, \mathcal{M}_{k-2}, \ldots$

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^T \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \sum_{\ell=1}^{k} \frac{\Theta_2 s_k^2 \left(1-\rho_M\right)^{k-\ell} \rho_M}{n} \sum_{i=1}^{n} \mathbb{E} \left\|\nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right)\right\|^2$$

Use property of $\{s_k\}$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^T \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\leq \sum_{k=0}^{T-1} \sum_{\ell=1}^{k} \frac{\Theta_2 s_\ell^2 \left(1 - \frac{\rho_M}{2}\right)^{k-\ell} \rho_M}{n} \sum_{i=1}^{n} \mathbb{E} \left\|\nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right)\right\|^2$$

$$\sum_{k=0}^{T-1} \sum_{\ell=1}^{k} \frac{\Theta_2 s_k^2 \left(1 - \rho_M\right)^{k-\ell} \rho_M}{n} \sum_{i=1}^{n} \mathbb{E} \left\|\nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right)\right\|^2$$

Use property of $\{s_k\}$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^T \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \sum_{\ell=1}^{k} \frac{\Theta_2 s_\ell^2 \left(1 - \frac{\rho_M}{2}\right)^{k-\ell} \rho_M}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right) \right\|^2$$

Use inequality $\sum_{k=1}^T \sum_{\ell=1}^k (1-\delta)^{k-\ell} \sigma_\ell \leq \frac{1}{\delta} \sum_{k=1}^T \sigma_k$, if $\delta \in (0,1]$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \sum_{\ell=1}^{k} \frac{\Theta_2 s_\ell^2 \left(1 - \frac{\rho_M}{2}\right)^{k-\ell} \rho_M}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{\ell+1}\right) - \nabla f_i\left(x_\ell\right) \right\|^2$$

$$\leq \sum_{k=0}^{T-1} \frac{2\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right) \right\|^2$$

Use inequality $\sum_{k=1}^{T} \sum_{\ell=1}^{k} (1-\delta)^{k-\ell} \sigma_\ell \leq \frac{1}{\delta} \sum_{k=1}^{T} \sigma_k$, if $\delta \in (0, 1]$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \frac{2\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right) \right\|^2$$

To make $D_f(y_k, x_k+1)$ and $\left\|x_k - y_k\right\|^2$ appear, use $\left\|a-c\right\|^2 \leq \left\|a-b\right\|^2 + \left\|b-c\right\|^2$ (triangular $+$ quadratic-arithmetic inequality).

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\leq \sum_{k=0}^{T-1} \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(y_k\right)\right\|^2$$

$$\sum_{k=0}^{T-1} \frac{2\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(x_k\right)\right\|^2$$

$$+ \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(y_k\right) - \nabla f_i\left(x_k\right)\right\|^2$$

To make $D_f(y_k, x_k + 1)$ and $\|x_k - y_k\|^2$ appear, use $\|a - c\|^2 \leq \|a - b\|^2 + \|b - c\|^2$ (triangular + quadratic-arithmetic inequality).

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^{T} \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$\sum_{k=0}^{T-1} \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(y_k\right)\right\|^2$$

$$+\frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(y_k\right) - \nabla f_i\left(x_k\right)\right\|^2$$

Apply Lipschitz continuity of ∇f_i and $\left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 \leq 2LD_f(y, x)$.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^T \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M\rho_F}$.

$$\sum_{k=0}^{T-1} \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(x_{k+1}\right) - \nabla f_i\left(y_k\right)\right\|^2$$

$$\leq 8 \sum_{k=0}^{T-1} \Theta_2 L s_k^2 \mathbb{E} D_f\left(y_k, x_{k+1}\right)$$

$$+ \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\left\|\nabla f_i\left(y_k\right) - \nabla f_i\left(x_k\right)\right\|^2$$

$$\sum_{k=0}^{T-1} + 4\Theta_2 L^2 s_k^2 \mathbb{E}\left\|x_k - y_k\right\|^2$$

Apply $\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2LD_f(y,x)$ and Lipschitz continuity of ∇f_i.

Bound on MSE

Let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$ and let $\{s_k\}_{k=1}^T \geq_0$ be any sequence s.t. $s_k^2(1-\rho) \leq s_{k-1}^2(1-\frac{\rho}{2})$. Let $\Theta_2 = \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}$.

$$8\sum_{k=0}^{T-1} \Theta_2 L s_k^2 \mathbb{E} D_f\left(y_k, x_{k+1}\right)$$

$$\sum_{k=0}^{T-1} + 4\Theta_2 L^2 s_k^2 \mathbb{E}\left\|x_k - y_k\right\|^2$$

Bias is bounded similarly.

Finite sum stochastic convex optimization

Parameters of usual estimators and rates under the accelerated meta theorem of this paper, up to constants (and complexity of a near optimal algorithm (Katyusha)):

unbiased biased

MSEB

	Full	aSVRG	aSAGA	aSARAH	aSARGE	Katyusha
M_1	0	$O(n/b^2)$	$O(n/b^2)$	$O(1)$	$O(1/n)$	-
M_2	0	0	0	0	$O(1/n^2)$	-
ρ_M	1	$O(b/n)$	$O(b/n)$	$O(1/n)$	$O(b/n)$	-
ρ_B	1	1	1	$O(1/n)$	$O(b/n)$	-
ρ_F	1	1	1	1	$O(b/n)$	-
cvx	$\frac{n}{\sqrt{\varepsilon}}$	$\frac{n}{\sqrt{\varepsilon}}$	$\frac{n}{\sqrt{\varepsilon}}$	$\frac{n^2}{\sqrt{\varepsilon}}$	$\frac{n^2}{\sqrt{\varepsilon}}$	$n \cancel{\log 1/\varepsilon} + \sqrt{n/\varepsilon}$
$\frac{\text{st. cvx}}{\log 1/\varepsilon}$	$n\sqrt{\kappa}$	$n^{2/3}\sqrt{\kappa}$	$n^{2/3}\sqrt{\kappa} + n$	$n^2\sqrt{\kappa}$	$n^2\sqrt{\kappa}$	$n + \sqrt{n\kappa}$